A Theory of Shape Identification

Pablo Musé, Frédéric Sur, Frédéric Cao, Jose Luis Lisani, Jean-Michel Morel

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A Theory of Shape Identification

P. Musé , F. Sur , F. Cao , J.L. Lisani , J.M. Morel

N°5766

Novembre 2005

Systèmes cognitifs

ISSN 0249-6399

INRIA
RENNES

A Theory of Shape Identification

P. Musé* , F. Sur† , F. Cao‡ , J.L. Lisani§ , J.M. Morel¶

Systèmes cognitifs
Projets Vista

Rapport de recherche n°5766 — Novembre 2005 — 181 pages

Abstract: What are shapes? Although shape recognition has long been one of the most important topics in computer vision, it is still hard to answer this question in a few words. In this book, shapes are defined as curves that can be recognized up to a given group of transformations. This definition, though tautological in appearance, has a very strong implication: how is it possible to identify two curves as being one and the same? The adopted point of view in this book is that no *a priori* model is necessary to take such a decision. A perceptual principle, the Helmholtz principle, will be the cornerstone of the decision. It asserts that two shapes should be identified if the probability that their resemblance may be due to chance is very small. Not only this principle may be useful in this identification step, but it will be also used throughout the complete system that will be presented: from the extraction of meaningful pieces of curves in digital images, to the grouping of invariant parts of shapes.

Key-words: Shape recognition, computer vision, Helmholtz principle, *a contrario* detection.

(Résumé : tsvp)

* CMLA, ENS-Cachan, muse@cmla.ens-cachan.fr
† CNRS/LORIA, sur@loria.fr
‡ INRIA/IRISA, fcao@irisa.fr
§ Dpt. de Ciències Matemàtiques et Informàtica, Universitat de les Illes Balears, joseluis.lisani@uib.es
¶ CMLA, ENS-Cachan, morel@cmla.ens-cachan.fr

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : 02 99 84 71 00 - International : +33 2 99 84 71 00
Télécopie : 02 99 84 71 71 - International : +33 2 99 84 71 71

Une théorie de la reconnaissance des formes

Résumé : Qu'est-ce qu'une forme ? Bien que la reconnaissance des formes ait été un des sujets majeurs de la vision par ordinateur, il n'y a pas de réponse simple à cette question. Dans ce livre, on définira les formes comme des courbes pouvant être reconnues, à un groupe de transformations géométriques près. La tautologie n'est qu'apparente : l'enjeu est bien de montrer qu'on peut prouver que deux courbes doivent êtres considérées comme identiques. Le point de vue adopté est qu'on peut se passer de modèle de formes. Un principe perceptuel, dit Principe de Helmholtz, sera le pilier central de cette théorie. Il stipule que deux formes doivent être considérées comme une seule et même forme, si la probabilité que leur ressemblance puisse être due au hasard est très faible. Ce principe sera utilisé à toutes les étapes du système qu'on présentera : non seulement à l'étape de la comparaison proprement dite, mais aussi à l'étape d'extraction de courbes pertinentes dans les images numériques, ainsi qu'au groupement de morceaux de formes, en entités cohérentes, les formes.

Mots-clé : Reconnaissance des formes, vision par ordinateur, principe de Helmholtz, détection *a contrario*.

Contents

Chapter 1

Introduction

1.1 A single principle

Digital images started to exist as a scientific object in the seventies of the past century. The emerging science dealing with digital images is called *Computer Vision*. It can be viewed as a realm of perception theory. It deals, however, with a much more affordable object than percepts. Indeed, digital images are just sampled real or vectorial functions defined on a part of the plane, usually a rectangle. They are accessible to all kinds of numerical, geometric and statistical measurements. In addition, the results of artificial perception algorithms can be confronted to human perception. This confrontation is both advantageous and dangerous: Experimental results may easily be misinterpreted by a visual exploration. In particular, these results will, still for long, look disappointing when matched with our perception. Now, it is the aim of Computer Vision to give, wherever possible, a mathematical definition of visual perception.
In a recent work by Desolneux, Moisan and Morel [47], a general mathematical principle, the so called Helmholtz principle, was proposed as a way to define all visual percepts as "large deviations from randomness".
According to this thesis one can compute detection thresholds deciding whether a given perceptual structure is present or not in each digital image. Several applications of this principle were developed by these authors and others for the detection of alignments [47], boundaries [48, 31], clusters [50], good continuation [28], vanishing points [2] and depth information recovery [123].
The mentioned authors make an extensive use of a computed function, the so called *number of false alarms (NFA)* of a perception. The NFA of a perception is the expected number of times this perception could have arisen "just by chance", by a casual arrangement of the background. An observed configuration in an image is a perception if and only if its NFA is much smaller than 1. Experimental evidence has confirmed that for human percepts, the NFA is actually much smaller than 1, typically less than 10^{-n} where n ranges from 10 to 100 and more [49].
So theory and experiments give a mathematical and experimental basis to the existence of sure percepts. Their existence had been conjectured for a long time by phenomenology, but without quantitative evidence.

The idea that visual objects are unlikely to arise in the background as perceptions goes back to Helmholtz [80]. Now, this principle could not be tested until images became digital and therefore accessible to computational experiments.
Before the above mentioned works, there had been several attempts to define percepts as exceptional events. C. V. Stewart [155] tried to detect planes in a cloud of points by what he called the MINPRAN method. He computed "the probability that MINPRAN will hallucinate a fit where there is none." This probability was computed under the *a contrario* assumption that the points were randomly distributed.
D. Lowe [111] proposed a detection framework based on the computation of accidental occurrence.

> *"In other words, we can shift our attention from finding properties with high prior expectations to those that are sufficiently constrained to be detectable among a realistic distribution of acciden-*

> *tals.[...] Even when we do not know the ultimate interpretation for some grouping and therefore its particular* a priori *expectation, we can judge it to be significant based on the non-accidentalness criteria."*

In the same spirit, Grimson and Huttenlocher [72, 73] proposed to compute shape recognition thresholds from a null model viewed as *"the conspiracy of random"*.

There is a common sense objection to the application of Helmhholtz principle to shape recognition. Watching the sky, one can often see castles, cats and dogs in the clouds. Humans have a high capacity for hallucinating familiar generic shapes such as faces in rich visual environments. This fact indicates that the Helmholtz principle is not adapted to all sorts of shapes.

The situation is, however, quite different when one talks about more specific, iconic shapes, such as letters, logos and in general all solid shapes. One may see faces in the sky, but certainly not this or this particular face. It is expected that any complex enough solid shape will be recognizable in the Helmholtz sense: No random arrangement would be able to reproduce it accurately.

The aim of these lecture notes is to prove that many sorts of shapes indeed are recognizable, or rather *identifiable*. All the above mentioned authors had remarked that the application of the Helmholtz method does not require any *a priori* model of the object to be built. One can be contented with what statisticians call a *background model*, or *a null model*. In the case of shape recognition, background is to be taken *à la lettre*: By the Helmholtz principle, a shape is conspicuous if and only if it could not be generated by the image background on which it is perceived.

The mathematics are quite simple. Let S and S' be two shapes observed in two different images and which happen to be close. Assume their Hausdorff distance is $\delta = d(S, S')$. Assume we know enough of the background model. More precisely, we are able to compute the probability $\Pr(S, \delta) = \Pr(d(S, \Sigma) \leqslant \delta)$ that some shape in the background, Σ be as similar to S as S' is. If this probability is very small, one would deduce that S' is not looking similar to S just by chance. Then S and S' can be *identified as the same shape*.

In a realistic setting, digital images contain thousands of significant *shape elements* that constitute their shape contents. (A precise numerical meaning to the notion of *shape element* will be given in this book.) The control of the number of wrong matches involves the probability of wrong match, but also the number of comparisons which are performed between the shape elements in two images.

DEFINITION 1.1 *Let $\mathcal{I}$ and $\mathcal{I}'$ be two images and N, N' the number of shape elements in each. Let S and S' be two shape elements extracted from I and I' respectively, lying at distance δ. We call number of false alarms of the match between S and S' the number*

$$NFA(S, S') = N \cdot N' \cdot \Pr(d(S, \Sigma) \leqslant \delta).$$

If $NFA(S, S')$ is much smaller than 1, one deduces that S' could not look like S just by chance and concludes that S and S' are, or have, *the same shape*.

There is an important phenomenological consequence: one can define shapes without the use of any empirical knowledge. By definition a shape is any part of an image which has been identified (in the sense of low NFA) at least once in another image.

From the empirical viewpoint, there are two kinds of shapes. First, any solid physical object can be photographed under many views and illumination conditions. If, by using the above definition, two snapshots of the same physical object happen to contain recognizable shape elements, one may say that the object itself is identifiable. These shape elements will constitute the object signature.

Second, notice that human activity builds all kinds of standardized objects. By the same method, two different standard objects can be identified, if they stem from the same industrial process. A similar remark applies to the very numerous iconic planar shapes generated by human visual communication, in particular the letters and logos.

In the experimental parts of this book, we shall study the identifiability of several such iconic shapes: the lower the NFA's they generate at a given Hausdorff distance, the more recognizable they are.

As a consequence of the present study, one should be able to define solid shapes as equivalence classes of recognized pairs, without any reference to any empirical knowledge or "ground truth". Thus, one should demonstrate the existence of, say, the Coca-Cola logo just by the fact that a certain group of shape elements appears in several images with very low NFA for all pairwise comparisons. Experiments will compare several snapshots of the same painting or poster, various images extracted from the same movie, or various logos of the same firm. The aim is, in all cases to single out and group in clusters all shape elements common to both images. Conversely, the same method should give a negative answer when two images have no shape in common.

From the mathematical and numerical viewpoint, the main challenge in the whole study is the accurate computation of small number of false alarms (NFA). This requires the computation of very small probabilities. Now, small probabilities cannot be measured from a shape database. A probabilistic model of the set of all possible shapes should therefore be built. Such a realistic experimental *background model* should be made of a large and representative set of digital images of all kinds. Unfortunately, there is no way to make a realistic probabilistic model of a large set of images. It looks as hopeless as to build a global "model of the world". Even if such a model were available, one would still face the challenge of computing accurately the probability of very rare events in this world model.

Fortunately enough, it is possible to overcome, or rather to go around these two obstacles. The only information needed is the probability for a background shape to be very close to a given query shape. This probability, which can be very small, cannot be learned from any database. By a geometric independence argument, this probability will be split into a product of much larger probabilities; these probabilities instead become observable in a small image database.

1.2 General overview

The above considerations on identification should not shadow some important aspects of shape modelling. Four main tasks must actually be performed properly on digital images to realize a shape identification: *extraction, encoding, recognition* and *grouping*.

1.2.1 Extraction

The first task is to define the shape elements which will be compared. Indeed, images are not compared globally, but detail to detail and up to several geometric and photometric perturbations which can alter them drastically. In the huge amount of raw information contained in a digital image, one therefore has to define the invariant features which will become shape characteristics. This part of the programme will be accomplished by a careful translation of several psychophysical invariance laws in mathematical and numerical terms. Following the Gestalt invariance laws proposed by Wertheimer, Attneave or Kanizsa [165, 13, 91], Chap. 2 develops an axiomatic point of view. The *shape elements* involved in shape recognition must be

- invariant to contrast changes;
- independent of the viewpoint, and therefore invariant by a subgroup of the projective group;
- insensitive to the noise inherent to any digital image;
- robust to partial occlusions, and therefore local enough.

The invariance requirements lead to the conclusion that shape elements must be obtained from "pieces of image level lines, adequately smoothed by an affine invariant smoothing process and contrasted enough". Chapter 3 deals with this extraction and displays many experiments. The numerical challenge is to extract as few level lines as possible from an image, with no loss in shape contents. At this stage, level lines are photometric invariants, but their geometric invariance is not achieved. We address the affine smoothing only briefly in

Sect. 4.3. This section explains why the robustness to noise and the invariance properties single out the affine curve scale space as the best multiscale representation of a planar curve.

1.2.2 Encoding

The second task is the invariant geometric encoding of the pieces of level lines. This is a tricky geometric computational issue which is treated in Part II. It describes procedures to locally encode the level lines to obtain *shape elements*. Such a local encoding is necessary to be robust with respect to partial occlusions.
This part involves some computational choices. There may be different representations fitting the same invariance properties and working as well. The local encoding basically consists in the choice of local frames on digital curves. Chap. 4 describes a way to compute stable directions on curves. These directions are then used in Chap. 5 to extract similarity or affine invariant shape elements.

1.2.3 Recognition

The third task and the main object of these notes is *identification*. This step is crucial, and is usually the Achilles' heel of shape recognition methods. So Part III is fundamental, short though it is. It aims at answering the question "are two given shape elements meaningfully alike"? The probabilistic modelling of the background model is performed in Chap. 6. It is tested by numerous experiments in Chap. 7.

1.2.4 Grouping

Decomposing shapes into local shape elements makes a recognition method robust to partial occlusions. A shape as a whole can be defined as a set of shape elements in a particular geometric configuration. The construction of these sets is the object of Part IV. Very interestingly, the problem can be formulated in terms of data point clustering. Clustering is one of the main techniques of "Pattern Recognition". To keep up with one of the main concerns here, automatic decisions, Chap. 8 focuses on the problems of cluster validation and of stopping rules in hierarchical clustering methods. These rules are applied to the grouping of shape elements in Chap. 9. Many examples on real images are given in Chap. 10.

In Appendix A are gathered some adds, particularly some short tutorials on the references used to write these lectures. Elements of comparison between the methods presented here and more classical ones are also developed. The keynotes can be read independently from the main text.

Acknowledgements The authors are indebted to their collaborators for many remarks and corrections, and more particularly to Andrés Almansa and Yann Gousseau. We chose not to give implementation details on the algorithms used for the experiments. They are all implemented in the public software MegaWave elaborated by Jacques Froment and Lionel Moisan. The research which led to the development of the present theory was mainly developed at the Centre de Mathématiques et Leurs Applications, ENS Cachan and CNRS, the Universitat de les Illes Balears and IRISA, Rennes. It was partially financed during the past six years by the Centre National d'Etudes Spatiales, the Centre National de la Recherche Scientifique, the Office of Naval research under grant N00014-97-1-0839 and the Ministère de la Recherche (project ISII-RNRT). Many thanks to Bernard Rougé and Wen Masters for their interest and constant support.

Part I

Extracting shape elements from images

Chapter 2

Shape psychophysics

2.1 What is a shape?

The concept of shape is wide ranging. The same remark holds, even if we restrict ourselves to the geometric visual shapes which are the object of this work. No matter what definition is adopted, one can hardly imagine a shape concept which does not primarily involve the identification problem. Following this point of view, phenomenologists [13, 117] conceive a shape as a subset of an image, digital or perceptual, endowed with some qualities permitting its recognition. Such a perceptual object is called a *planar shape*. Making this remark into a definition, one can call shape any part of an image which can be recognized in another image.

Humans reliably recognize shapes undergoing a wide range of transformations and perturbations that will be described later. This leads us to the more accurate definition:

DEFINITION 2.1 (GENERAL DEFINITION OF SHAPE) *Let $\mathcal{I}$ and $\mathcal{I}'$ be any two different images in W. We call* shape in *I, I' any part common to $\mathcal{I}$ and $\mathcal{I}'$* modulo *a class of invariance. Let W be a set of reference images ("the world of possible images"). We call* shape in *W any shape common to a pair of images of W.*

From a practical viewpoint, this definition is still too general: "common part" has not yet a precise meaning, and the class of invariance has not been specified.

2.1.1 Invariant properties of shape recognition and a well adapted image representation

In order to find the shape invariance classes, it is enough to give a rough typology of the transformations that affect images but not our recognition of shapes therein. Following Lisani *et al.* [107], the main classes of perturbations which do not affect recognition are:

1. **Changes in the color and luminance scales (contrast changes).** According to gestaltists Attneave [13] and Wertheimer [165], shape perception is independent of the grey level scale or of the measured colors.

   > *"The concentration of information in contours is illustrated by the remarkable similar appearance of objects alike in contour and different otherwise. The "same" triangle, for example, may be either white on black or green on white. Even more impressive is the familiar fact that an artist's sketch, in which lines are substituted for sharp color gradients, may constitute a readily identifiable representation of a person or thing."* Attneave, 1954.

   > *"I stand at the window and see a house, trees, sky.*
   > *Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees. It is impossible to achieve "327" as such. And yet even though such droll calculation were possible and implied, say, for the house 120, the trees 90, the sky 117 – I should at least have this arrangement and division of the total, and not, say, 127 and 100 and 100; or 150 and 177."* Wertheimer, 1923.

We refer to Figure 2.1 designed by E. H. Adelson for a striking illustration of illumination invariance.

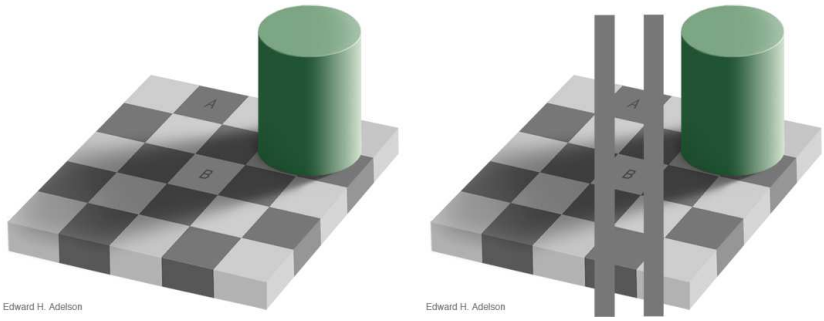


Figure 2.1: Contrast change invariance. In the left hand image, the A and B squares have exactly the same grey level. This incredible fact is easily checked in the right hand image where A and B are linked by two rectangles with the same grey level. This experiment due to E.H. Adelson illustrates the unreliability of brightness perception and the invariance of shape recognition with respect to illumination changes. (Courtesy E.H. Adelson, http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html)

2. **Occlusions and background modification.** Shape recognition can also be performed in spite of occlusion and varying background, as shown in Figure 2.2. The phenomenology of occlusion was thoroughly studied by Kanizsa [91] and his school. Kanizsa argues that occlusion is always present in every day's vision: most objects are partially hidden by other ones. Human perception must therefore be able to recognize partial shapes. Conversely, if a shape occludes a background, its recognition is invariant to changes in the background. This relative independence of shape recognition from its background is known in perception psychology as the *figure-background problem.* It was thoroughly studied by Rubin [144]. The figure-background problem is another aspect of the occlusion problem. A shape is superimposed to a background, which can be made of various objects: how to extract, to single out, the shape from that clutter? This leads to a dilemma: Extract the shape and then recognize it or, conversely, extract it *because* it has been recognized?

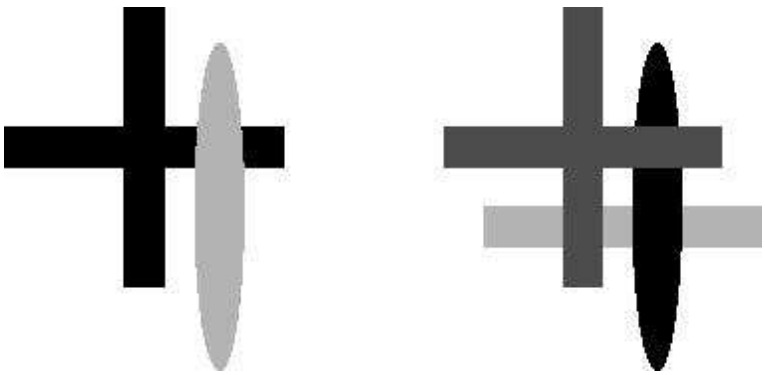


Figure 2.2: Left: According to the theory of G. Kanizsa and his school, shapes can be recognized even when they undergo several occlusions. Our perception is trained to recognize shapes which are only seeable in part. Here the occluded cross can be easily recovered. Right: the *figure-background* problem. Our perception is adapted to recover a figure on the foreground, independently from the background

3. **The classical noise and blur**, inherent to any perception task and to any image generated according to Shannon's theory.

4. **Geometrical distortions or deformations.** The effects of perspective are deeply incorporated in human perception. Humans can recognize objects and shapes under perspective distortion, as long as perspective is not too strong. Recognition is also invariant to elastic deformations, always within some limits.

The previous four invariant properties fix some rules or requirements a good image representation should comply with. Therefore, it is necessary to formulate a mathematical model of each of them, so as to derive a well adapted image representation.

1. (a) **The local contrast invariance requirement** A digital image is usually defined as a function $u(x)$, where $u(x)$ represents the grey level or luminance at x. The first task is to extract from the image a topological information fairly independent from the varying and unknown contrast change function of the optical and/or biological apparatus. One can model such a contrast change function as any continuous increasing function g from $\mathbb{R}^+$ to $\mathbb{R}^+$. The real datum corresponding to the observed u could be as well any image $g(u)$. This simple argument leads to select the level sets of the image [150], or its set of level lines, as a complete contrast invariant image description [33]. If u is of class say C^1, then the level lines are the connected components of $u^{-1}(\lambda)$, which are C^1 curves for almost every $\lambda \in \mathbb{R}$.

   (b) **The concentration of information requirement.** Somewhat in contradiction with this contrast invariance principle, many authors assert, like Attneave [13] , that *"Information is concentrated along contours (i.e., regions where color changes abruptly)"*. One can argue that not all the level lines are really needed to have a complete description in terms of perception. Some of them are due to noise or to small, hardly noticeable, changes in illumination. Thus, it makes sense to select only the most contrasted level lines, that is to say, those along which the gradient of u is large enough. Such a selection (and any other) is not invariant to any contrast change, since it explicitly uses the gradient value. However, it can be shown that it is invariant to affine global contrast changes. Besides, it is probably the most stable selection, in the sense that these lines will not vary significantly when "not too strong" contrast changes are applied. A simplification of the level lines tree can be performed by using the method proposed by Desolneux *et al.* [48], which greatly reduces the number of level lines, while preserving the main structures in the image. Figure 2.3 shows an example of such level lines selection (see caption for details).

2. **The occlusion and figure-background requirements.** Even the best adapted choice of level lines is not totally suited to describe image parts. Indeed, when a shape A partially occludes a shape B, the level lines of the resulting image are a concatenation of pieces of the level lines belonging to A and to B. This is shown with a very simple example in Figure 2.4. Even if a shape is not occluded, but simply occludes its own background, there may be no level line surrounding the whole shape, as shown in figure 2.5. These remarks show that the whole level lines, do not provide all "shape candidates". In order to overcome this obstacle, a segmentation of level lines into their parts belonging to different objects is needed. This seems to create another hen-and-egg problem. A segmentation of level lines would require that the different objects are known in advance! The only way to overcome this ignorance is to segment the level lines in small enough pieces. It is then expected that most of these pieces will locally coincide with the boundary of a single object.

3. **The smoothing requirement.** If "common parts" in images subject to noise are still recognized, this means that shape information has not been affected by noise. In that sense, noise can be viewed as introducing details which are much too fine (in relation to the essential shape information) to be perceptually relevant, in terms of recognition. Let us quote Attneave (ibid., 1954):

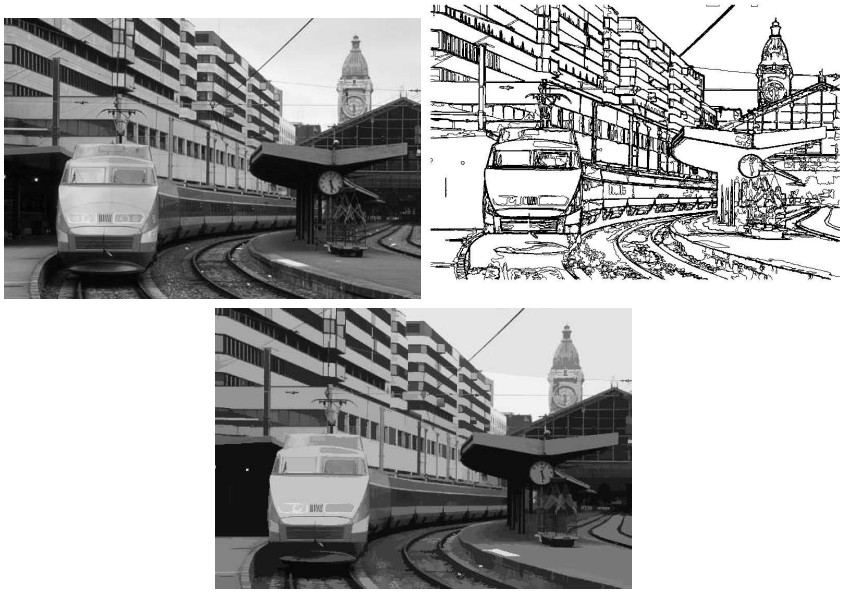


Figure 2.3: Original image on the top left (83, 759 level lines). Top right: meaningful boundaries (883 level lines). Bottom: reconstruction from the meaningful boundaries. Only 883 boundaries remain, while the structure of the image is preserved and perceptual loss is very weak

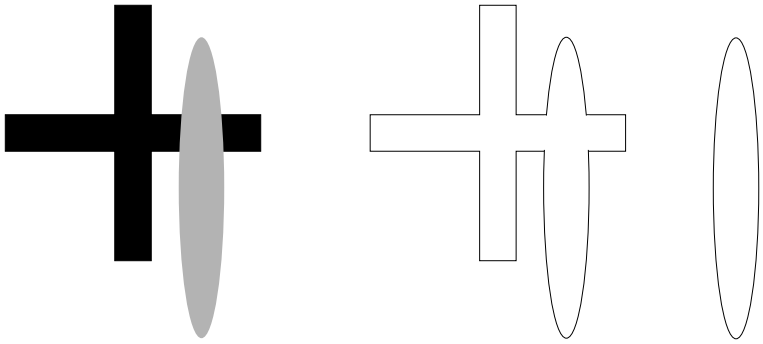


Figure 2.4: Left: oval occluding a cross, right: the level lines of the resulting image. While the boundary of the oval can be recovered as a full level line, the boundary of the cross concatenates with the oval's boundary. Thus recognition cannot be based on complete level lines, but it can still be based on pieces of level lines

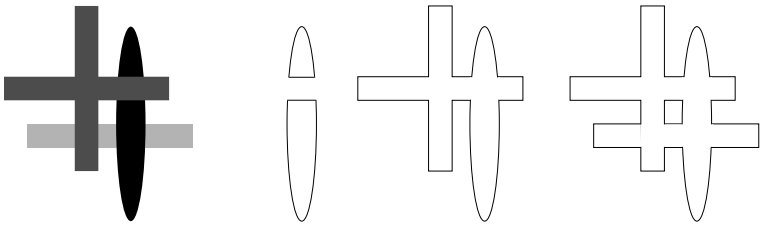


Figure 2.5: Left: Cross on a background with an oval occluding a rectangle. The cross is wholly in view. All the same, its shape does not appear as a level line because of the background. As in Figure 2.4, one sees that the level lines must be broken into pieces to get clues of each single shape

> *"It appears, then, that when some portion of the visual field contains a quantity of information grossly in excess of the observer's perceptual capacity, he treats those components of information which do not have redundant representation somewhat as a statistician treats "error variance", averaging out particulars and abstracting certain statistical homogeneities."*

Hence, a correct image representation, which does not get lost in textural details, asks for a previous smoothing. This fact is illustrated by Figure 2.6. The object on the right was obtained by smoothing the one on the left. Both objects differ in their small details. Nevertheless, most people would recognize a "black disk" on both sides.

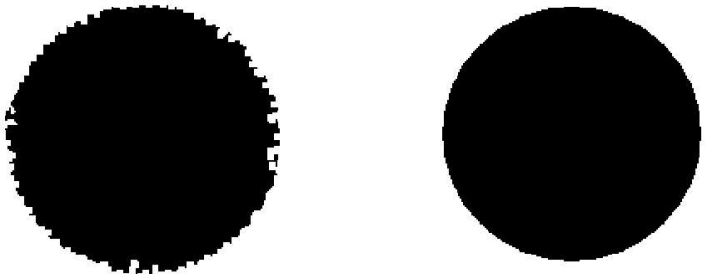


Figure 2.6: One can immediately see that both objects are disks, with approximately the same size. The second one is obtained from the first by the affine curvature equation [5]. The main ideas behind such a curvature equation was anticipated by Attneave, who proposed to smooth silhouettes by blurring and then enhancing the resulting image to get a smooth silhouette: "somewhat as if the photograph of the object were blurred and then printed on high-contrast paper"

4. **Geometric invariance requirements.** Image representations (a set of meaningful level lines, for instance) have to be invariant to weak projective transforms. Allowing invariance to any projective transform does not make sense, since one cannot recognize shapes under strong perspectives. Besides, it can be shown that all planar curves within a large class can be mapped arbitrarily close to a circle by projective transforms. This result was reported by Åström in [12], where it is also shown that given a finite set of m Jordan curves $\mathcal{C}_1, \ldots, \mathcal{C}_m$, one can find a Jordan curve $\mathcal{C}$ and m projective transforms $p_1, \ldots, p_m$, such that $p_i(\mathcal{C})$ is arbitrarily close to $\mathcal{C}_i$, for all $i \in \{1, \ldots, m\}$. Hence, in general, schemes based on projective normalization of Jordan curves are not possible. Another argument against general projective invariance is that, despite some interesting attempts [60], there is no practical way to define a projective invariant local smoothing. From this viewpoint, affine invariant smoothing is the "best" compromise [5].

Deriving an image representation

To end with this section, let us summarize the four invariance requirements, and the constraints they impose to an image representation based on shape information. The *local contrast invariance* led us to define the image level lines as a contrast invariant representation. The *concentration of information requirement* leads to the selection of a set of *meaningful level lines* that are roughly the level lines which are long and contrasted enough (a precise definition will be given in Chap. 3). It follows, from the *occlusion and figure-background requirements*, that *small pieces* of meaningful level lines are to be considered. Combining the *smoothing requirement* and the *geometric invariance requirement* pieces of meaningful level lines imply that level lines should be smoothed by a an affine invariant transform. This leads us to the following definition of *shape elements*, as the elementary structures likely to be recognized under all mentioned perturbations.

DEFINITION 2.2 *We call* shape element *of an image any piece of well contrasted level line of the image, affine invariantly smoothed. This piece of level line is not considered by itself, but rather by its equivalence class under all affine transforms of the plane.*

The next chapter deals with the “well contrasted" requirement. Part II will define an adequate numerical representation of the affine equivalence classes of shape elements.

Chapter 3

Extracting meaningful curves from images

he set of level lines of an image (isophotes), or topographic map, provides a complete, contrast invariant, representation of an image. This representation is well suited for shape analysis and recognition, since it contains the geometrical information of images. It has a tree structure for inclusion, reflecting the intuitive shape inclusions. It is therefore sound to use level lines as the proper shape representatives for any shape recognition algorithm. Now, some complexity issues have to be handled first: The number of level lines in 8 bits encoded images is typically 10^5. Most of them are very small lines that are due to noise. Moreover, level lines in texture usually do not have proper shapes. So in order to make the computational burden of shape recognition algorithms affordable, it is useful and even necessary to select the "best" level lines, namely the ones that are the most susceptible to correspond to image contours. In this chapter, a method to select the most meaningful level lines (boundaries of level sets) from an image is presented. This extraction can be based on the Helmholtz Principle. By this method, the number of encoded level lines is reduced by a factor 100, without any signifcant loss of shape contents. The starting point is an algorithm first published by Desolneux, Moisan and Morel [48]. A mathematical interpretation of the model is given, explaining why some pieces of curve are over detected. A method to correct this flaw is then proposed, as well as a multiscale approach that makes the method more robust to noise. A more local variant of the algorithm is also introduced, taking local contrast variations into account.

3.1 The level lines tree, or topographic map

A grey level digital image u_d is a function defined in a rectangular grid, that takes values in a finite set, typically integer values between 0 and 255. Such a datum must be interpolated to obtain a grid independent representation. According to Shannon's theory, this interpolation must be band limited and is therefore analytical. Simpler spline interpolation methods classically can provide interpolates of arbitrary regularity class C^k for any $k \in \mathbb{N}$. All of these interpolates yield for $k \geqslant 1$ level lines with a simple topological structure.

THEOREM 3.1 *Let u be a C^1 image in $\mathbb{R}^2$. Then, for almost all $\lambda \in \mathbb{R}$ and for all compact domain $K \subset \mathbb{R}^2$, the set $K \cap \partial(u^{-1}[\lambda, +\infty))$ is a finite set of C^1 Jordan curves. These curves, called the level lines of u are either closed or meet the boundary of K at exactly two points.*

This is an easy consequence of Sard's Theorem and the implicit functions theorem: It is enough that λ avoids the singular values of u. In this case level lines are connected subsets of the isophotes of u, that is the set of points where u assumes a constant value. In the following, level lines of u are defined as all Jordan curves obtained by taking all level lines at all non critical levels, as indicated by the above theorem. The geometrical information in images can then be reduced to the *level lines*. The *topographic map* of an image, defined as the collection of all of its level lines, gives a complete representation of an image and satisfies two main properties:

- It is invariant with respect to contrast changes. Indeed, if g is an increasing function from $\mathbb{R}$ to $\mathbb{R}$, then u and $g(u)$ have the same level lines (up to a set of levels with measure 0).

- It is a hierarchical representation: level lines at different levels do not meet, since the image is a continuous function. So the topographic map can be embedded in an inclusion tree structure.

In the following, a bilinear interpolation shall actually be used, (order 1 spline.) The bilinear interpolates are Lipschitz but not C^1. Now, it is easily checked that almost all level lines still are Jordan curves and piecewise C^1. So the above mentioned tree inclusion structure is still valid. Among the possible interpolations, the bilinear interpolation presents two advantages: It is the most local of all classical continuous interpolations and does not create new extrema in the image, contrarily to Shannon's interpolation that creates ringing effects. There is no need to compute level lines at too many levels. It is in practice enough to take all levels $n+\frac{1}{2}$ where n goes from 0 to 255. This choice minimizes grid effects, as illustrated in Figure 3.1.

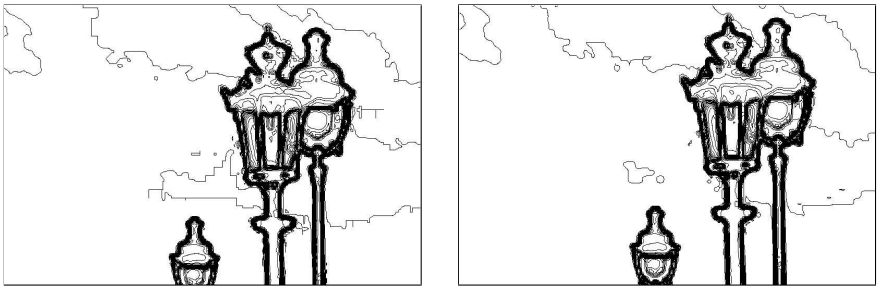


Figure 3.1: Left: level lines from the piecewise bilinear interpolated image. The quantization step for the grey levels is 10, starting from 10. Some grid effect ("pixelization") can be seen. Right: level lines from the piecewise bilinear interpolated image, with a grey level quantization step of 10, starting at grey level 0.5. These level lines do not suffer from pixelization effects

3.2 Meaningful boundaries

This section addresses the problem of selecting the most contrasted level lines in an image, as selected in the psychophysical theory of Attneave [13]. Now, this selection involves at least two measurements, namely the length of the level line and its contrast. Intuitively, a short very contrasted level line should be kept, but so should be a less contrasted but longer one. The correct weighting out of these two criteria is the object of the present chapter. The *a contrario* method of Desolneux et al. [48] is adopted, according to which a level line is a meaningful boundary if it could not appear in noise; now this definition needs a careful discussion. The *a contrario model* thus involved is far from obvious, and the practical challenge here is to extract as few level lines as possible without missing any shape of interest in a digital image.

3.2.1 Contrasted boundaries

Let $u : \mathbb{R}^2 \to \mathbb{R}$ be a differentiable grey level image.[1] Let us assume that a measure of contrast is given at each point, which is taken here to be the norm of the gradient. What are the level lines of u along which the contrast assumes unexpectedly high values? This sentence refers to the *a contrario* model in which contrast values would be randomly distributed on the level line, with a law identical to the contrast law in the whole image. This contrast law can be approximated by the empirical histogram. So it is assumed that the gradient norm is distributed following the law of the positive random variable X defined by

$$\forall \mu > 0, \quad P(X > \mu) = \frac{\#\{x \in \Gamma,\, |Du(x)| > \mu\}}{\#\{x \in \Gamma,\, |Du(x)| > 0\}}, \tag{3.1}$$

[1] If u is a bilinearly interpolated image, then it is Lipschitz continuous and piecewise C^1. Thus its gradient is a L^∞ function, defined everywhere except on the mesh linking the center of pixels, which is a negligible set.

where the symbol # denotes the cardinality of a set, Γ the finite sampling grid and $|Du|$ is computed by a finite difference approximation. In the following, the inverse repartition function is denoted by

$$H_c(\mu) = P(|Du| > \mu).$$

In [48], Desolneux et al. proposed the following definition.

DEFINITION 3.1 ([48]) *Let N_{ll} be the number of level lines of u. A level line C with length l is an ε-meaningful boundary if*

$$NFA(C) \equiv N_{ll} H_c(\min_{x \in C} |Du(x)|)^{l/2} < \varepsilon, \tag{3.2}$$

This quantity is called the number of false alarms (NFA) of C.

In (3.2), the NFA is the product of the number of level lines by the probability that a random curve Γ containing $\frac{l}{2}$ independent samples has its contrast larger than $\min_{x \in C} |Du(x)|$ everywhere, when the contrast values on curve samples are assumed to be mutually independent. If the NFA is very small, this means that this assumption is certainly not valid, leading to an *a contrario* detection. Remark that meaningful boundaries are not invariant with respect to general contrast change. Indeed, let $g : \mathbb{R} \to \mathbb{R}$ be a C^1 function and $v = g(u)$. Let C be a level line of u with level λ and $\mu = \min_{x \in C} |Du(x)|$. Then C is a level line of v with level $g(\lambda)$. However, $|Dv| = |g'(u)||Du|$. It is impossible to estimate the distribution of $|Dv|$ from the distribution of $|Du|$, since there is no relation between $g'(u)$ and $|Du|$.
A trivial, though interesting exception, is affine contrast change. If $g(s) = as + b$, then $|Dv| = |a| \cdot |Du|$. Hence, if $a \neq 0$, the inverse repartion function of the norm of $|Dv|$ is $H'(\mu) = H_c\left(\frac{\mu}{|a|}\right)$. Therefore

$$H'(\min_{x \in C} |Dv(x)|) = H'(|a| \min_{x \in C} |Du(x)|) = H_c(\min_{x \in C} |Du(x)|),$$

and the number of false alarm of C in v is the same as in u. This proves the following result.

LEMMA 3.1 *Meaningful boundaries are invariant to affine contrast changes.*

3.2.2 Maximal boundaries

As remarked by Desolneux et al. [48], meaningful boundaries usually appear in parallel groups, because of the blur inherent to all well-sampled images according to Shannon's theory. In order to eliminate the redundancy of contrasted boundaries, these authors use the inclusion tree structure described in Sect. 3.1. Using the standard terminology of trees (nodes, branches, leaves), let us simply recall that the nodes of the tree are the level lines of the image. The ordering is defined by the inclusion. This means that a Jordan level line C_1 is the mother of another one C_2 if it surrounds it, and there is no other one surrounding C_2 and surrounded by C_1. The leaves of the tree are then the level lines which do not surround any other one.

DEFINITION 3.2 ([129]) *A monotone section of a tree of level lines is a part of a branch such that each line has a unique daughter and where the grey level is monotone (no contrast reversal).*
A maximal monotone section is a monotone section which is not strictly included in another one.

DEFINITION 3.3 ([48]) *A meaningful boundary is maximal meaningful if it has a minimal NFA in a maximal monotone section of the tree of meaningful level lines.*

Figure 3.2 shows how negligible the information loss is when representing an image by its maximal meaningful boundaries. They represent roughly one hundredth of all level lines.
Since maximal meaningful boundaries inherit the tree structure of the tree of level lines , they can be used to reconstruct an image (see Fig. 3.3.)

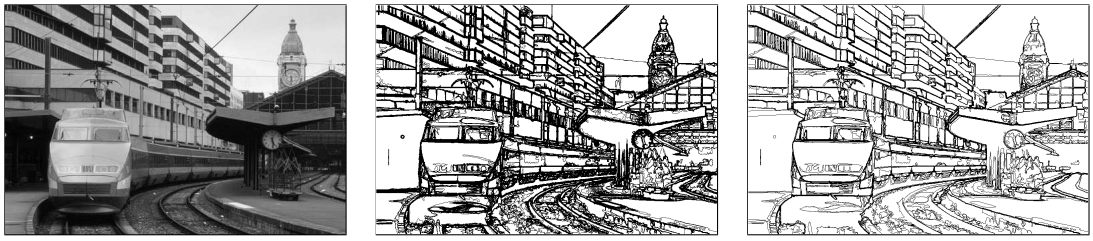


Figure 3.2: Maximal meaningful boundaries. 1. Original image, 83,759 level lines 2. All meaningful boundaries: 11,505 detections. 3. Maximal meaningful boundaries. Only 883 boundaries remain, almost no detail is lost

Figure 3.3: Original image on the left (99,829 level lines). Right: reconstruction from the 394 maximal meaningful boundaries. The grey level is constant and equal to the average image grey level in each connected component of the complementary of these level lines. Notice how the main shape features are preserved, while textures are removed. This simplification and reconstruction algorithm is obtained from a pruning of the tree of level lines. Salembier and Serra [145] call such operators *connected operators*.

3.2.3 A mathematical justification of meaningful contrasted boundaries

In this section, the precise interpretation of Def. 3.1 is given. It follows that the definition does not prevent meaningful boundaries from containing parts with low contrast. From this, a simple cleaning rule will be derived, aiming at removing these parts.

Interpretation of the number of false alarms

The following classical lemma will be used several times within this book.

LEMMA 3.2 *Let X be a real random variable. Let $F(x) = P(X \leqslant x)$ the repartition function of X. Then, for all $t \in (0,1)$*

$$P(F(X) < t) \leqslant t.$$

In the same way, let $H(x) = P(X \geqslant x)$. Then for all $t \in [0,1]$,

$$P(H(X) < t) \leqslant t.$$

Proof: Let us define the pseudo-inverse

$$F^{-1}(t) = \inf\{s,\ F(s) \geqslant t\}. \tag{3.3}$$

Because of the convention in its definition, F is right-continuous. Hence

$$F \circ F^{-1}(t) \geqslant t.$$

Moreover, for all $x \in \mathbb{R}$,

$$F(x) < t \Leftrightarrow x < F^{-1}(t). \tag{3.4}$$

Indeed, let us first assume that $F(x) < t$. If $x \geqslant F^{-1}(t)$, then $F(x) \geqslant F \circ F^{-1}(t) \geqslant t$ (because F is nondecreasing), which is a contradiction. Conversely, let us assume $x < F^{-1}(t)$. Then, $F(x) \geqslant t$ would contradict the definition of $F^{-1}(t)$. This proves the equivalence. Hence,

$$\begin{aligned} P(F(X) < t) &= P(X < F^{-1}(t)) \quad \text{by (3.4)} \\ &= P(\exists y,\ y < F^{-1}(t),\ X \leqslant y) \\ &= \sup_{y < F^{-1}(t)} F(y) \\ &\leqslant t \quad \text{again by (3.4).} \end{aligned}$$

The third equality is a basic convergence theorem of measure theory. Let us note that the last inequality is large, because of the passage to the limit. The second part of the lemma is proved in the same way by introducing $H^{-1}(t) = \sup\{s,\ H(s) \geqslant t\}$. The proof is left to the reader. □ Let us remark that if F is continuous and increasing, F^{-1} is really the inverse of F and the Lemma then yields an equality, and means that $F(X)$ is a uniform variable in $(0,1)$.

Let us assume that X is a real random variable described by the inverse repartition function $H(\mu) = P(X \geqslant \mu)$. Assume that u is a random image such that the values of $|Du|$ at each point in the sample grid are independent, and follow the same law as X. Let now E be a set of random curves (C_i) in u such that $\#E$ (the cardinality of E) is independent of each C_i. For each i, let $\mu_i = \min_{x \in C_i} |Du(x)|$. Let us also assume that L_i independent points can be chosen on C_i. The curves C_i can be thought of as random walks with independent increments but since a finite number of samples are selected on each curve, the law of the C_i does not really matter. Let us finally assume that L_i is independent from the pixels crossed by C_i. Such a random model is called *a contrario* or *background* model. For instance, some straight lines in white noise with a given length

distribution satisfy these hypotheses.
A curve C_i is said to be ε-meaningful if

$$NFA(C_i) = \#E \cdot H(\mu_i)^{L_i} < \varepsilon.$$

Remark: In digital images, the independence of the values of $|Du|$ is sound only if the points are far enough from each other. In practice, the minimal distance will be taken equal to 2, since a 2×2 finite difference scheme is used to compute the image gradient. The following proposition justifies Def. 3.1.

PROPOSITION 3.1 *The expected number of ε-meaningful curves in a random set E of random curves is smaller than ε.*

Proof: Let us denote by X_i the binary random variable equal to 1 if C_i is meaningful and to 0 else. Let also $N = \#E$. Let us denote by $\mathbb{E}(X)$ the expectation of a random variable X in the a contrario model. Then

$$\mathbb{E}\left(\sum_{i=1}^{N} X_i\right) = \mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{N} X_i | N\right)\right).$$

It is assumed that N is independent from the curves. Thus, conditionally to $N = n$, the law of $\sum_{i=1}^{N} X_i$ is the law of $\sum_{i=1}^{n} Y_i$, where Y_i is a binary variable equal to 1 if $nH(\mu_i)^{L_i} < \varepsilon$ and 0 else. By linearity of the expectation,

$$\mathbb{E}\left(\sum_{i=1}^{N} X_i | N = n\right) = \mathbb{E}\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} \mathbb{E}(Y_i).$$

Since Y_i is a Bernoulli variable, $\mathbb{E}(Y_i) = P(Y_i = 1) = P(nH(\mu_i)^{L_i} < \varepsilon) = \sum_{l=0}^{\infty} P(nH(\mu_i)^{L_i} < \varepsilon | L_i = l) P(L_i = l)$. Again, it is assumed that L_i is independent of the gradient distribution in the image. Thus conditionally to $L_i = l$, the law of $nH(\mu_i)^{L_i}$ is the law of $nH(\mu_i)^l$. Let us finally denote by $(\alpha_1, \cdots, \alpha_l)$ the l (independent) random values of $|Du|$ along C_i. Then,

$$\begin{aligned}
P\left(nH(\mu_i)^l < \varepsilon\right) &= P\left(H(\min_{1 \leqslant k \leqslant l} \alpha_k) < \left(\frac{\varepsilon}{n}\right)^{1/l}\right) \\
&= P\left(\max_{1 \leqslant k \leqslant l} H(\alpha_k) < \left(\frac{\varepsilon}{n}\right)^{1/l}\right) \text{ since } H \text{ is nonincreasing} \\
&= \prod_{k=1}^{l} P\left(H(\alpha_k) < \left(\frac{\varepsilon}{n}\right)^{1/l}\right) \text{ by independence} \\
&\leqslant \frac{\varepsilon}{n} \text{ from Lemma 3.2.}
\end{aligned}$$

This term does not depend upon l. Thus

$$\sum_{l=0}^{\infty} P(nH(\mu_i)^{L_i} < \varepsilon | L_i = l) P(L_i = l) \leqslant \frac{\varepsilon}{n} \sum_{l=0}^{\infty} P(L_i = l) = \frac{\varepsilon}{n}.$$

Hence,

$$\mathbb{E}\left(\sum_{i=1}^{N} X_i | N = n\right) \leqslant \varepsilon.$$

This finally implies $\mathbb{E}\left(\sum_{i=1}^{N} X_i\right) \leqslant \varepsilon$, what means exactly that the expected number of ε-meaningful curves is less than ε. □

In this proposition, it is not assumed *a priori* that the C_i are level lines of u. Indeed, in this case, it cannot certainly be asserted that the length (number of independent points) of the curve is independent from the values of the gradient along the curve.

Cleaning-up meaningful boundaries

Proposition 3.1 asserts that if a curve is a meaningful boundary, then it cannot be *entirely* generated in white noise (up to ε false detections on the average). On the other hand, can it be guaranteed that no part of a meaningful boundary is contained in noise? Or, for a given meaningful boundary, is it possible to give an upper bound of the size of the part of the boundary that is likely to be contained in noise (i.e. a non-edge region)? To answer this question, let us introduce the *a* posteriori length distribution

$$P(L \geqslant l | \min_{x \in C} |Du(x)| \geqslant \mu), \tag{3.5}$$

where the image model is a white noise, C is a level line in this image and L its length. The computation of this probability needs the *a priori* distribution of L, through $P(L \geqslant l)$. This last distribution is not explicitly known, but since the image model is white noise, it can be empirically estimated. For $l \leqslant 1000$ (to give an order of magnitude), the number of lines whose length is larger than l is still quite large (for images of size about 500×500), and the distribution is assumed to be correctly estimated for such length. (See Fig. 3.4.) For higher values, there are too few level lines. By using Bayes' rule,

$$P(L \geqslant l | \min_{x \in C} |Du(x)| \geqslant \mu) = \frac{\sum_{k=l}^{\infty} P(\min_{x \in C} |Du(x)| \geqslant \mu | L = k) P(L = k)}{\sum_{k=1}^{\infty} P(\min_{x \in C} |Du(x)| \geqslant \mu | L = k) P(L = k)}.$$

(The denominator is nothing but $P(|Du| \geqslant \mu)$). By the a contrario assumption (independence of the gradient

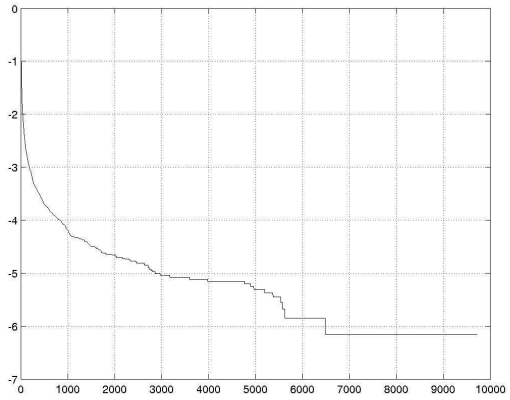


Figure 3.4: Log10 of the inverse repartition function of length of level lines in a white noise image. The average length is about 3.5, meaning that most level lines enclose a single pixel

along curves),

$$p_\mu(l) \equiv P(L \geqslant l | \min_{x \in C} |Du(x)| \geqslant \mu) = \frac{\sum_{k=l}^{\infty} H_c(\mu)^k P(L = k)}{\sum_{k=1}^{\infty} H_c(\mu)^k P(L = k)}. \tag{3.6}$$

Remark that in (3.6), the gradient value appears on the right-hand side only through its probability. Let us now consider an image u with N_{ll} (quantized) level lines. Let us also denote by N_l the number of all possible sampled subcurves of these level lines. (If the curves are closed with L_i independent points, then $N_l = \sum_{i=1}^{Nu} L_i^2$.)

LEMMA 3.3 *For all $l > 0$, the map $\mu \mapsto p_m u(l)$ is non increasing.*

Proof: Set $P(L = k) = a_k$ and $H_c(\mu) = x$. Then $a_k \geqslant 0$, $\sum_k a_k = 1$, $x \in (0, 1)$, and

$$\frac{1}{p_\mu(l)} = 1 + \frac{\sum_{k=1}^{l-1} a_k x^k}{\sum_{k=l}^{\infty} a_k x^k}.$$

Let $y = H_c(\nu)$ for $\nu \geqslant 0$. Then $\nu \leqslant \mu \Rightarrow y \geqslant x$. The sign of $\frac{1}{p_\mu(l)} - \frac{1}{p_\nu(l)}$ is also the sign of

$$N(x,y) = \left(\sum_{k=1}^{l-1} a_k x^k\right)\left(\sum_{k=l}^{\infty} a_k y^k\right) - \left(\sum_{k=1}^{l-1} a_k y^k\right)\left(\sum_{k=l}^{\infty} a_k x^k\right).$$

Since $y \geqslant x$, and $a_k \geqslant 0$,

$$\sum_{k=l}^{\infty} a_k y^k \geqslant \left(\frac{y}{x}\right)^l \sum_{k=l}^{\infty} a_k x^k,$$

and

$$\sum_{k=1}^{l-1} a_k y^k \leqslant \left(\frac{y}{x}\right)^l \sum_{k=l}^{\infty} a_k x^k.$$

Plugging this into the definition of $N(x,y)$ yields $N(x,y) \geqslant 0$ for $x \leqslant y$. Hence, $p_\mu(l)$ is nonincreasing with respect to μ. □

Assume that C is a piece of level line with L independent points, contained in a non-edge part, described by the noise model. The problem is to estimate the probability that L is larger than $l > 0$, knowing that $|Du| \geqslant \mu$. This is exactly $p_\mu(l)$, the probability defined in (3.6). As in Prop. 3.1, it can be proved that $N_l \cdot p_\mu(l)$ is an upper bound of the expected number of *pieces* of lines of length larger than l with gradient norm larger than μ. For a fixed μ, let l be such that $N_l \cdot p_\mu(l) \leqslant \varepsilon$. Then, on the average, less than ε pieces of level line are observed with a length larger than l and a gradient norm everywhere larger than μ.

Let us now choose a small value of μ and make the assumption that a point x with a gradient norm less than μ is located in noise. Pieces of curve of length less than l and containing x, may appear ε times in average. If they are simply removed, then all remaining points belong to a piece of curve with length larger than l and with gradient norm larger than μ, which cannot be due to chance.
This yields a clean-up algorithm for boundary detection:

1. Detect meaningful boundaries.
2. For a fixed $\mu > 0$, let $\mathcal{L}(\mu) = \inf\{l,\ N_l \cdot p_\mu(l) < \varepsilon\}$.
3. For any meaningful boundary, remove every subcurve of length $\mathcal{L}(\mu)$ containing a point where $|Du| \leqslant \mu$.

A new parameter, μ, has been introduced. When μ gets larger, $\mathcal{L}(\mu)$ decreases, so that the clean-up procedure removes more numerous but smaller pieces of curves.

The choice of μ can be determined by applicative considerations. Detected edges may be used for different purposes, for instance shape recognition or image matching. If $|Du|$ is less than 1, then the distance between two level lines with level differing by 1 is larger than 1. This mean that boundaries do not have an accurate position. Choosing to eliminate pieces of curves with a gradient larger than $\mu = 1$ for all images is therefore not restrictive in shape recognition applications. It is also easily checked in experiments that for $\mu = 1$, $\mathcal{L}(\mu)$ is less than a few hundreds, which is compatible with the empirical estimation of the *a priori* length distribution. Figure 3.5 shows an example of the clean-up procedure.

3.3 Multiscale meaningful boundaries

As previously noted above, the contrast measure is an approximation of the gradient by finite differences. More precisely, Desolneux et al. use the following scheme:

$$\frac{\partial u}{\partial x} \simeq u_x(i,j) = \frac{1}{2}(u(i+1,j) + u(i+1,j+1) - u(i,j) - u(i,j+1)), \tag{3.7}$$

$$\frac{\partial u}{\partial y} \simeq u_y(i,j) = \frac{1}{2}(u(i,j+1) + u(i+1,j+1) - u(i,j) - u(i+1,j)). \tag{3.8}$$

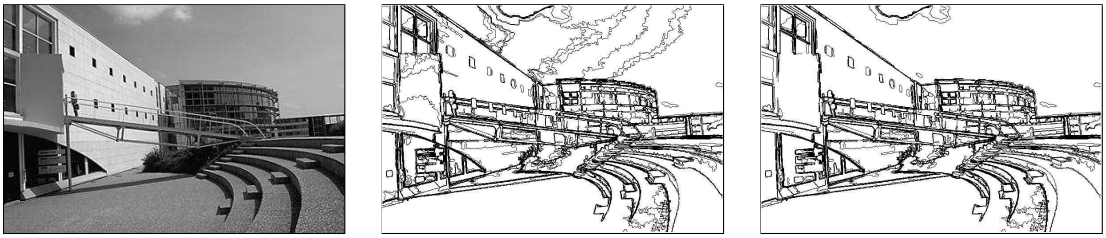


Figure 3.5: Meaningful boundary clean-up. On the left the original image. In the middle, the meaningful boundaries with local histograms, see Sect. 3.4. Boundaries are found in the sky. They are detected, because the gradient in the sky is regular due to the smoothly changing illumination. The gradient value is about 0.2 in the sky, but the curves are so long that they are detected. This does not contradict our detection principle: such curves are indeed exceptional in noise, since it is very unlikely that the gradient does not attain an even smaller value on such a long curve. What is actually contradicted is our assumption that these exceptional curves do correspond to edges, no matter how small the contrast is. This assumption indeed implies that one is able to distinguish arbitrary grey level changes. This is perceptually not true. On the right, result after the clean-up procedure with a gradient threshold equal to 1

Using a 2×2 scheme is coherent with the application of Helmholtz principle: points at distance 2 have independent values of contrast in white noise. On the other hand, this value is sensitive to noise. Smoothing the image before computing the gradient would partly remove noise but would also introduce local dependencies between pixels. This would make the *a contrario* model false in smoothed white noise, and false detections could be expected. Nevertheless, the a contrario models still applies if the image is downsampled at a lower rate, in conformity with Shannon's sampling theory.

More precisely, the following algorithm is applied. Consider a set of N_s dyadic scales $\{1, 2, ..., 2^{N_s-1}\}$. For any level line C, let us denote by C^s the curve $\frac{C}{2^s}$, obtained by scaling C by a factor 2^{-s}. Let also H^s denote the empirical contrast distribution of u^s, where u^s is obtained by downsampling u by a factor 2^s in agreement with Shannon's theory. (That is to say, downsampling follows an adequate smoothing, for instance convolution with a prolate function.)

1. Compute the quantized level lines of u.

2. For each level line C with l independent points in u, compute μ^s, the minimal value of $|Du^s|$ over all pixels crossed by C^s. Let

$$NFA(C) = N_s \cdot N_{ll} \min_{s \in \{0, \cdots, N_s - 1\}} (H^s(\mu^s))^{l/2^s}. \tag{3.9}$$

   A curve C is an ε-meaningful multiscale boundary if $NFA(C) < \varepsilon$.

Thus, a curve is meaningful if and only if there exists a scale such that it is $\frac{\varepsilon}{N_s}$ meaningful in the sense of the previous section. Roughly speaking, the N_s factor is the price to pay to have the right to test several different scales. It is clear that it only make sense to consider a small number of dyadic scales (say 3 or 4, since the side of usual digital image does not exceed much 2^{10} pixels). Since the detection depends on $\log \varepsilon$ [47, 48], considering $\frac{\varepsilon}{N_s}$-meaningful boundaries at each scale can eliminate only a few lines.

The expected number of detections in a white noise image is still under control.

COROLLARY 3.1 *With the same hypotheses as in Prop. 3.1, the expected number of ε-meaningful multiscale boundaries in a white noise image is less than ε.*

Indeed, downsampled images are still white noise images. Together with the linearity of expectation and the proof of Prop. 3.1, this yields the result.

Figures 3.6 and 3.7 show the result of this multiscale method on images with quantization noise and additive Gaussian noise. On Fig. 3.6, the shapes are not very sharp because of motion blur and transparency. Level lines following contours are very long since they surround several objects. Moreover, the background is nearly uniform. Thus the minimal value of contrast along long level lines is all the more sensitive to the gradient computation. The effect is also dramatic in the noisy image of Fig. 3.7 (Gaussian noise with standard deviation 30). Note also how the boundaries of the main objects still coincide with level lines, in spite of the very strong noise.

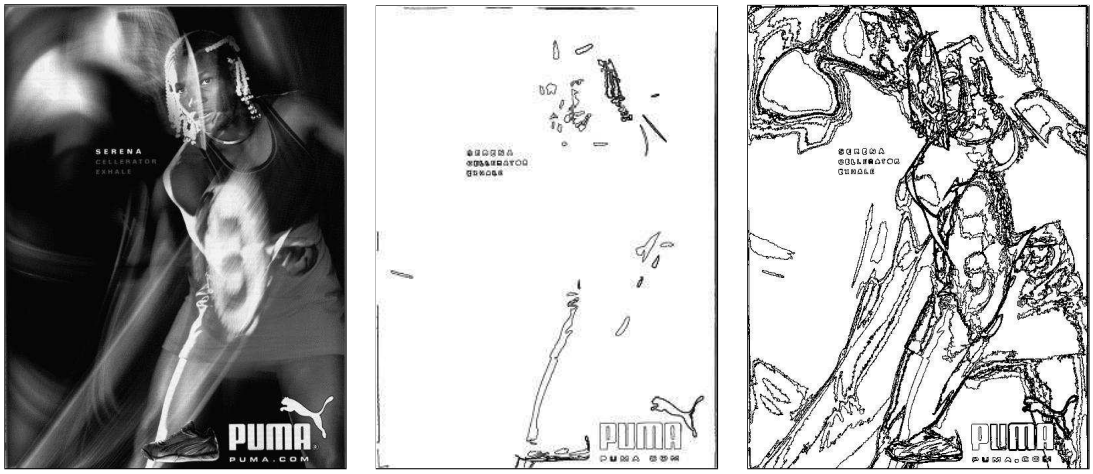


Figure 3.6: Influence of quantization noise on meaningful boundaries. On the left, the original image is coarsely quantized since it has a very low contrast. This leads to bad gradient estimation and a lot of missing detections (middle). Multiscale detection is less sensitive to quantization noise and leads to more correct detections (right)

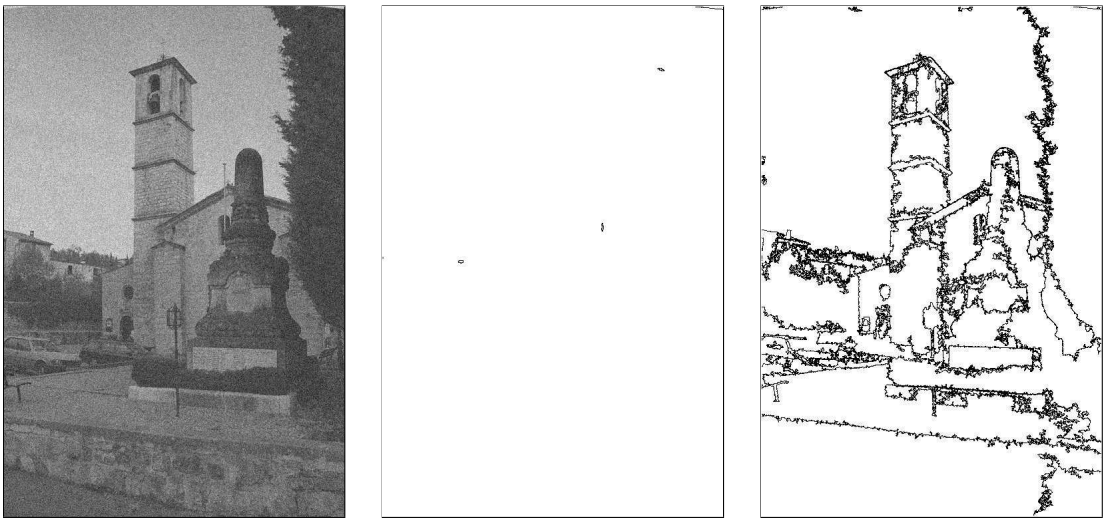


Figure 3.7: Multiscale meaningful boundaries and noise. Left: image of Fig. 3.9 with an additive white Gaussian noise of standard deviation 30. Middle: meaningful boundaries. Since noise dominates the gradient distribution, only six small level lines are detected. Right: multi-scale detection using four dyadic scales. Textures are not detected, meaning that noisy textures are in this case not different enough from noise to be detected. On the other hand, main structures remain. This allows one to empirically check the stability of the topographic map in spite of the important amount of noise

3.4 Local boundary detection

In the above *a contrario* model, the values of the gradient are random variables whose distribution is empirically estimated by using the histogram of the gradient in the image. The use of this global distribution yields the so-called "blue sky effect". Consider an image containing two parts: a contrasted or textured one (e.g. ground) and a smooth one (e.g. sky), see Fig. 3.9. There is an empirical overdetection in the ground, and an underdetection in the sky. Indeed, the sky only contributes with small values in the histogram. Thus the algorithm tends to detect any level line which is more contrasted than the sky. So nearly anything is detected in the ground. Conversely, the contrasted ground may make the detection more difficult for regions with a small contrast, like a sky with clouds. This is not quite coherent with human vision, which locally adapts the perception of contrast. This section addresses this local adaptivity to contrast, by modifying the meaningful boundary model. It describes the algorithm and shows experiments.

3.4.1 Algorithm

Let us assume that a closed boundary has been detected. It divides the image into two connected components: the interior and the exterior of the curve. Let us compute the empirical contrast distribution in each component. Meaningful boundaries are then detected, independently in each connected component. This procedure can now be recursively applied. Since the tree of level lines of a quantized image has a finite depth, it is clear that the detection procedure stops after a finite number of steps.
Two problems make things slightly more complicated. First, the order that is used to describe the image boundaries may have an influence. The most natural solution is simply to start with the most meaningful boundaries.

A second problem is purely computational and involves open boundaries, whose endpoints belong to the image border. They still cut the image into two connected components, that should be processed in the same way since there is no clear notion of interior and exterior. However, in order to make the tree structure unique, exactly one of these components is considered as the interior. Open boundaries are then "closed" following the shortest path along the border of the image. This choice is only an algorithmic one, and is arbitrary from a perceptual point of view [129]. To circumvent this lack of symmetry between both connected components, the detection is first applied to open boundaries, until no new open boundaries are detected. The procedure is then applied to closed boundaries.

Let us specify the algorithm coming from the above considerations.
Let us call R_0 the root boundary, that is the (non-meaningful) boundary containing all the image. If C is a boundary, its interior is denoted by $\operatorname{Int} C$.

1. Set $R \leftarrow R_0$. (Local root.)
2. Set $\mathcal{M}$, the set of already stored in R meaningful boundaries. Initially, $\mathcal{M}$ is empty.
3. Let $R' \leftarrow R \backslash \cup_{C \in \mathcal{M}} \operatorname{Int} C$.
4. Compute the histogram of $|Du|$ in R'.
5. Use this histogram and detect the maximal meaningful boundaries included in R'. Let us call total maximal boundaries, the meaningful boundaries C satisfying

$$\begin{cases} \operatorname{Int}(C') \subsetneq \operatorname{Int}(C) \Rightarrow NFA(C) < NFA(C') \\ \operatorname{Int}(C) \subset \operatorname{Int}(C') \Rightarrow NFA(C) \leqslant NFA(C'). \end{cases} \tag{3.10}$$

The set of total maximal boundaries is denoted by $\mathcal{N}$. In other words, the boundaries in $\mathcal{N}$ have an optimal NFA, since they are more meaningful than boundaries which contain them or in which they are contained. This assumption is stronger than the maximality defined in Sect. 3.2.2 since the NFA comparison is not restricted to monotone sections. The subtree with root equal to R that remains by

keeping only the boundaries in $\mathcal{N}$ has only two levels: the local root R, and $\mathcal{N}$. Since the interior of open boundaries is arbitrary, the detection of open and closed boundaries are not mixed. In practice, this means that if an open meaningful boundary C is detected, the definition of total maximal boundary (3.10) is only applied to open boundaries containing C or contained in C.

6. If $\mathcal{N} \neq \emptyset$, then new boundaries have been detected in the complementary of the already detected ones. Then,

   (a) Set $\mathcal{M} = \mathcal{M} \cup \mathcal{N}$. By construction, all the closed boundaries in $\mathcal{M}$ have disjoint interior.

   (b) return to step 3.

7. If $\mathcal{N} = \emptyset$, there are no new boundaries in the local root and in the complementary of the currently detected boundaries. The search is then resumed at lower levels of the tree, as follows. For any boundary $C \in \mathcal{M}$,

   (a) Store C.

   (b) Set $R \leftarrow C$, and $\mathcal{M} \leftarrow \emptyset$.

   (c) Return to step 3.

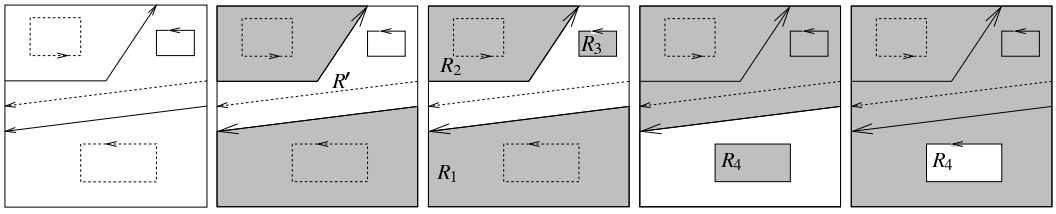


Figure 3.8: Example of local search of meaningful boundary. (a) the initial boundaries. They are oriented such that the tangent and the inner normal form a direct frame. The NFA of each boundary is computed. There are three total maximal boundaries (in solid line); two are open, one is closed. While some open curves are detected, the closed ones are skipped. (b) Compute the contrast histogram in the complementary set of the interior of the open detected boundaries and resume search in this part of the image, which is the region R'. The closed boundary is total maximal meaningful again. However, no new open boundaries are detected. Thus, this closed boundary is kept. (c) The search is resumed (with recomputed histogram) in the exterior (white part) of the detected boundaries, until new ones can no longer be found, which is the case on the figure. When this is over, compute the local contrast histogram in each region R_1, R_2, R_3 and look for boundaries inside them. (d) A (closed) total maximal boundary R_4 has been detected in R_1. Compute the local histogram in $R_1 \backslash R_4$ and detect boundaries. (e) Finally, scan for boundaries in R_4 with new local contrast histogram. Since nothing is detected, the output is the boundaries of R_1, R_2, R_3 and R_4

Remark: Each boundary may be tested more than once. Thus, the number of false alarms has to be multiplied by the maximal number of visits of a boundary, which is bounded by above by the depth of the level lines tree. In fact, each detected boundary often lies in the middle of the local root, and this divides the tree depth by 2. Thus the maximal number of visits of a boundary is of the order of the logarithm of the initial tree depth. In practice, it is always much smaller than 100.

3.4.2 Experiments on locally contrasted boundaries

Figure 3.9 shows the difference between the detection with a global contrast histogram and the updated local histogram. To give an idea of the magnitude of the number of false alarms, the boundary separating sky and the foreground has a NFA equal to 10^{-357}. This means that such a contrasted line in noise should be observed at most once every 10^{357} level lines of white noise images. The smaller boundaries around the opening on the top of the tower have NFAs about 10^{-10}.

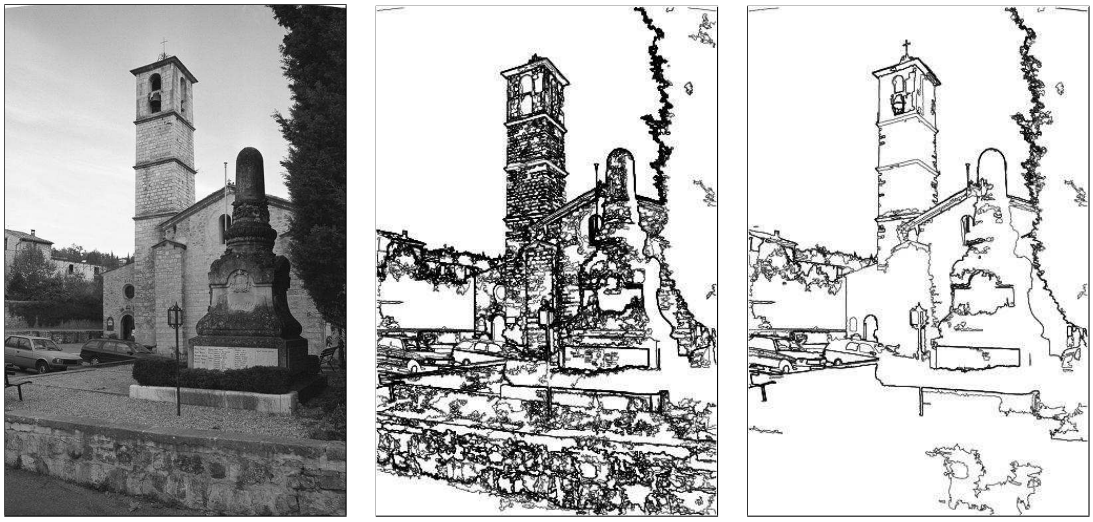


Figure 3.9: Influence of local contrast. From left to right: original image, maximal meaningful boundaries, local maximal meaningful boundaries. There are 280,000 boundaries in the initial image (for a grey level quantization step of 1), 652 in the second one and 193 in the last one. Texture is removed since local contrast (for instance) on the church tower is much more demanding than the global histogram. As the texture is uniform, no level line is a large deviation to the empirical local contrast, yielding no detection. This is very good for shape analysis where it is often desirable to distinguish texture from real shapes

The effect of local contrast in boundaries detection is twofold

1. textures are eliminated.
2. On the contrary, local contrast should make curves in low contrasted areas more detectable.

This was expected since, in textured regions (as on the tower), the local contrast values are larger than in the rest of the image. Thus, this increases the NFA of boundaries; most of them simply disappear in textured regions. This is a masking phenomenon in the terminology of Gestalt theory [91].
On the other hand, some lines are detected due to the illumination gradient (See Fig. 3.5 and 3.10). They can be due to the vicinity of the light source, or to the variation of the orientation of the surface of a three dimensional object with respect to the light source. Such lines do not correspond to the silhouette of physical objects. Nevertheless, it is reasonable to detect them as remarkable structures.

What is the impact of the preceding study for our scopes in shape recognition? It is well known that texture is strongly damaged by compression. Thus, the precise geometry of level lines in texture may depend very much on the image source (quality, compression rate etc...). Moreover, they are very complex, and will yield many encoded pieces of curves when the procedure of Chap. 5 is applied. The shape content of a texture is therefore both huge in quantity and unreliable. The computational cost to handle it may therefore be too high for some applications. Thus it may be useful to automatically remove contrasted regions corresponding to texture.
The argument above is reversed for stereo image registration or motion estimation. In this case, it is *a priori* known that the images under comparison are nearly the same image. The goal is to register them as best

as possible. In this application, textures generate many level lines which can be tracked and should not be eliminated.

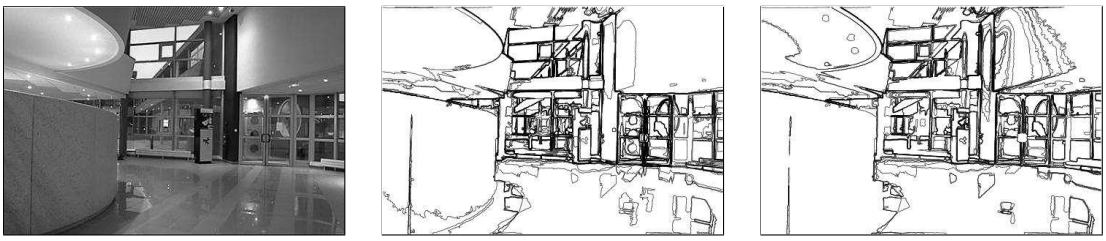


Figure 3.10: Illumination, local contrast and regularity. Left: original image. Middle: meaningful contrasted boundaries. Right: meaningful contrasted and smooth boundaries with local contrast. (See. Sect. A.1.2.) With contrast only, a single boundary appears on the right with the contrast due to illumination. If contrast is localized, then more boundaries are detected. If a regularity constraint is also considered, there are still more detections. These boundaries are very different from texture, and from noise since they are nearly convex and parallel at low scale.

3.5 Bibliographical notes

The presentation of this chapter mainly follows [31], which presents a mathematical discussion and major improvements of the boundary detection method proposed in [48]. In continuation, some genealogy for edge detection, level lines, level sets, and the topographic map will be given. In the keynote A.1.2.2 at the end of this book, a paragraph analyzes the relationship between maximal meaningful level lines and the well-known snakes, or active contours variational method. In particular, the role of the regularity of level lines is discussed in a still more sophisticated *a contrario* method.

3.5.1 Edge detection

(See also many more details and comparison in Section A.1.2.1.) It is a well known fact that shape information in images is concentrated along regions where color or grey level changes abruptly [13, 114]. Since Marr and Hildreth's seminal work on edge detection [115], the effort on extracting shape information from images has been mainly concentrated on local methods. Among these methods, which are commonly referred as edge detectors, Canny [25] and Canny-Deriche [45] filters are certainly the most widely used.

Classical edge detectors suffer from two problems. The first one is that they depend on (at least) two parameters, the threshold on the contrast and the degree of smoothing. Both are hard to estimate and are usually fixed manually. The second problem of these methods is that they detect points with an orientation, and not structures. These points have to be connected afterwards to build again curves. Such chaining algorithms involve further parameters.

3.5.2 Level lines and shapes

Following [107], in Chapter 1 it was asserted that the set of level lines of a digital image was a natural representation of its shape content, since it provides a geometric information invariant to contrast changes . Moreover, no chaining procedure is needed since level lines are already curves. This chapter has presented the bilinear level line tree proposed by Lisani *et al.* [108]. Whereas edge detectors usually fail near T-junctions (and additional treatments are necessary), there are several level lines at a junction (See Fig. 3.11) and [32]. This will be detailed in Sect.A.1.1.

Prior to the use of level lines, shape analysis was performed in Mathematical Morphology by associating with any image a family of binary images obtained by thresholding at all levels. This yields a complete

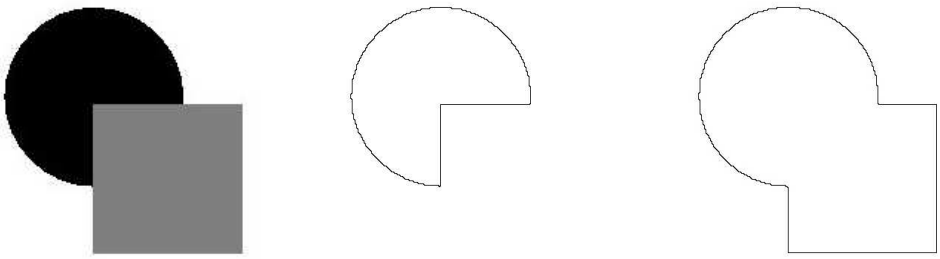


Figure 3.11: Level lines and T-junctions. Depending on the grey level configuration between objects and background, level lines may follow or not (as on the figure) the objects boundary. In any case, junctions appear where two level lines separate. Here, there are two kinds of level lines: Those surrounding the occluded circle and those following the boundary of the union of the circle and the square. These level lines are included in each other and do not meet, but are usually very close and not distinguishable along contrasted contours

representation of the image by its upper level sets. [116, 150]. The (upper) level set of a grey level image $u : \mathbb{R}^2 \to \mathbb{R}$ at the value λ is defined by

$$\chi_\lambda(u) = \{x \in \mathbb{R}^2, \quad u(x) \geqslant \lambda\}.$$

An image can be reconstructed from the whole family of its level sets, by

$$u(x) = \sup\{\lambda \in \mathbb{R}, \quad x \in \chi_\lambda(u)\}.$$

The *level lines* are obtainable by simply taking the boundaries of upper level sets. The tree structure of the topographic map has been extensively used to build an efficient computational representation of the level lines; see the algorithms of Monasse et al. [130, 15, 108]. An efficient region growing algorithm, the *Fast Level Set Transform* allows one to compute the tree of level lines for digital images (constant in each pixel) or bilinearly interpolated images [108]. The idea to consider the level lines of the bilinear interpolated image was also independently proposed in the so-called Digital Morse Theory [40].

3.5.3 Extraction of shapes from images

The extraction of shape elements is seldom addressed in the context of shape recognition. Most works on shape recognition assume that shapes are already extracted [65, 124, 142]. In Mokhtarian's approach [124, 125, 126], shapes are extracted by simply thresholding dark objects over a bright background, so get back to upper level sets. Their boundaries are level lines. Rothwell proposed a whole recognition system of flat objects on uniform background [142]. The shapes he treats are simply Jordan curves bounding objects. Rothwell's method builds the object boundaries by extracting edges using Canny's edge detector [26]. Canny's filter performs well in Rothwell's framework, where objects are well-contrasted over a uniform background. In general, this filter is not that efficient (see A.1) but in that particular easier case, one simply gets back level lines again!

Part II

Geometric encoding of shape elements

Chapter 4

Robust directions of shapes

his chapter deals with shape normalization, a method which associates with all shapes deduced from each other by an affine distortion a single, normalized, shape. A crucial ingredient for normalization is the computation of a small affine covariant set of robust straight lines associated with a shape. The set of all tangent lines of a shape has this covariance property, but it is too large. A very successful idea from the computational viewpoint is to involve bitangent lines, that is, lines tangent at two different points of a shape. If the shape has finitely many inflexion points, then it also has finitely many bitangent lines. So in Section 4.3, a well-established technique to smooth Jordan curves in such a way as to reduce their number of inflexion points will be briefly described, this smoothing being in addition affine invariant. Convex shapes have no bitangent lines and simple shapes have only very few bitangent lines. It has shown very useful in shape recognition to compute other robust straight lines associated with the shape. Flat parts of curves are informally defined as intervals of the curve along which the direction of the tangent line does not vary too much. For instance, large enough polygons show as many reliable flat parts as sides. A simple parameterless definition of flat parts is based on an *a contrario* model. As a formal definition, the flat part detection algorithm *extracts all flat parts of all level lines of an image which could not appear in white noise.* Of course, any part of any smooth curve can be considered as flat at some scale. Not so in an *a contrario* model: only pieces of curves which are *exceptionally* flat in the *a contrario* model can be defined as flat parts. In that way, one gets a relative, but parameterless definition of flatness.

4.1 Flat parts of level lines

The concept of "flatness" of a part of curve is measured in what follows by how much the curve turns on this part with regard to the direction given by the underlying chord (see figure 4.1).

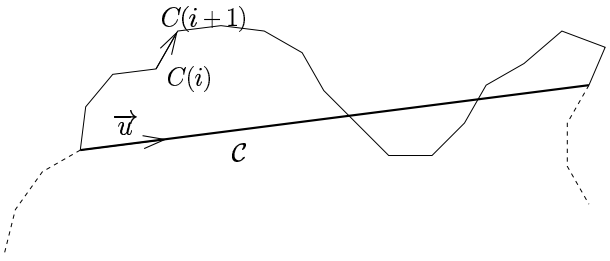


Figure 4.1: A piece of discrete curve with the underlying chord $\mathcal{C}$ (thick segment line)

Although flatness may look like a rather intuitive geometric concept, it is in fact quite complex. Our aim is to define a unique measurement, the "flatness" for very diverse phenomena: A long very oscillating curve may look flat seen at a distance. In another way, a short and very smooth curve can look locally very flat. One can therefore figure out that at least two parameters are involved in a flatness measurement, one measuring the

length of the flat part and another giving the amplitude of oscillations. Thus, the flatness definition problem can be viewed as the question of reducing two parameters to a more abstract one, the "flatness". The detection of flat parts of a curve should meet the following requirements:

- It should not detect just points around which the curve is flat, but the precise intervals on which the curve is straight.
- Long flat parts should be allowed to move farther from their underlying chord than short ones.
- The detection should be intrinsic to the curve, and not depend on other curves in the image.
- Detected flat parts should not overlap.
- Since the flat parts detection is generally the first step of a recognition algorithm, it deals with a huge amount of information. Therefore, computational complexity should be low.

4.1.1 Flat parts detection algorithm

Let us consider a chord from a given curve C: its endpoints delimitate a piece of curve of length l (measured in pixels). Since one would like to measure how much the piece turns with respect to the direction $\overrightarrow{u}$ given by the chord, let us define

$$\alpha = \max_{i \in \{0 \dots n-1\}} \left\{ \left| \text{angle}(\overrightarrow{C(s_i)C(s_{i+1})}, \overrightarrow{u}) \right| \right\},$$

where the discrete piece of curve is made of the n consecutive points $C(s_i)$.

Let us suppose that α is below some fixed threshold α^*. Following the discussion on independence in Sect. 3.3, let us consider that points at a geodesic distance (along the curve) larger than 2 are statistically independent. Thus, there are $l/2$ statistically independent segments of the type $(C(s_i), C(s_{i+1}))$ along a curve with length l. The probability of the event "$l/2$ statistically independent points on a piece of curve show a tangent line which makes an angle lower than α among all the pieces of curve for which $\alpha < \alpha^*$" is:

$$p(\alpha, l) = \left(\frac{\alpha}{\alpha^*} \right)^{l/2}.$$

Of course, the lower $p(\alpha, l)$, the more flat the piece of curve.

This straightforward computation is valid under the assumption that among all the pieces of curves such that $\alpha < \alpha^*$, α is uniformly distributed over $[0, \alpha^*]$, and independent at Nyquist distance 2. Flat parts are now defined as rare events with regard to this *a contrario* model.

For each piece of the curve for which $\alpha < \alpha^*$, the probability $p(\alpha, l)$ is computed. Only pieces such that $p(\alpha, l)$ is under a predetermined threshold p^* are kept (these parts are called "candidates"). Such pieces can of course overlap. So a selection has to be made among them in order to define the flat parts of the curves. A greedy algorithm will be used in practice: the piece of curve with the lowest p is marked as a "flat part", then all candidates that share a common part with this "best" flat part are eliminated. The process is iterated with the remaining candidates.

4.1.2 Reduction to a parameterless method

The computation of α clearly depends on the discretization. The curves which the proposed algorithm deals with are level lines of images. Their "natural" discretization is the pixel, that appears to be accurate enough in order to compute α.

The whole algorithm involves two thresholds. The first one, α^*, is not critical. Indeed, since one is interested in detecting flat parts, it is natural to *a priori* reject all pieces of curve for which α is upon a large threshold. We set $\alpha^* = 1$ radian once for all, which is not a strong constraint. Let us be more specific on this issue. A change of α^* multiplies all probabilities $p(\alpha, l)$ by a constant factor. Thus, the flatness measurement is just scaled and

the ordering maintained. Moreover, changing α^* also multiplies the threshold p^* by the same constant. Thus, there are not two parameters here, but just one, namely p^*. This last parameter will be eliminated, as it can be fixed in such a way that almost no flat part occurs in the level curves of a white noise.
Indeed, experimental evidence shows that $p^* = 10^{-3}$ is the maximum value for which only a few detections (on the average one) can be made on the level lines extracted from a white noise image, containing the same amount of level lines than a standard natural image. So the proposed algorithm satisfies the Helmholtz principle: almost no detection of flat parts in a white noise image.

4.1.3 The algorithm

The flat parts detection algorithm is summarized in what follows.

Let us consider a Jordan curve on which flat parts are searched for.
Part I: candidate identification.
For each chord of the curve with length $10, 20, 30, \ldots, 180, 200$, and then an exponential progression[1]:

1. Compute the maximum angle α between the chord and the piece of curve delimited by both ends of the chord. If n denotes the number of independent points $C(s_i)$ on this piece of discrete curve:

$$\alpha = \max_{i \in \{1 \ldots n-1\}} \left\{ \left| \text{angle}(\overrightarrow{C(s_i)C(s_{i+1})}, \overrightarrow{u}) \right| \right\}.$$

2. If $\alpha > 1$ rad, *a priori* reject the piece; otherwise compute $p(\alpha, l) = \left(\frac{\alpha}{\alpha^*}\right)^{l/2} = \alpha^{l/2}$, where l is the length of the considered piece of curve.

3. If $p(\alpha, l) > p^* = 10^{-3}$, reject the piece.

Part II: greedy algorithm

1. Keep the candidate for which $\alpha^{l/2}$ is minimal, mark it as *flat part*, and discard it from the list of candidates.

2. Reject all candidates that meet this "best" candidate.

3. Iterate until no candidate is available anymore.

4.1.4 Some properties of the detected flat parts

The condition defining the candidates ($\alpha^{l/2} < p^*$) is not a real constraint for long curves. For example, if $p^* = 10^{-3}$ and $l = 200$, all curve parts such that $\alpha < 0.97$ are accepted as candidates. Nevertheless, long pieces of curves often show subparts with a lower probability and a greedy algorithm will therefore prefer them. In the case of circles, however, this does not occur. Let us compute the arcs of circle which will be marked as flat parts. Figure 4.2 illustrates the following computations.

PROPOSITION 4.1 *A circle of radius R has flat parts if and only if $R \geqslant -e \log(p^*)$.*
In such a case, the length of the detected flat parts is $L = 2R \sin(1/e)$.

[1]There is a complexity issue here. "All" chords are not tested, but a subsample of them so that the algorithm does not waste too much time for long curves. The only consequence of this discretization procedure is that long straight lines (in practice, lines whose length is larger than 100 pixels) could be split into two pieces (see Figure 4.14 for an example). This is not an important drawback, since the goal is to use flat parts as robust directions.

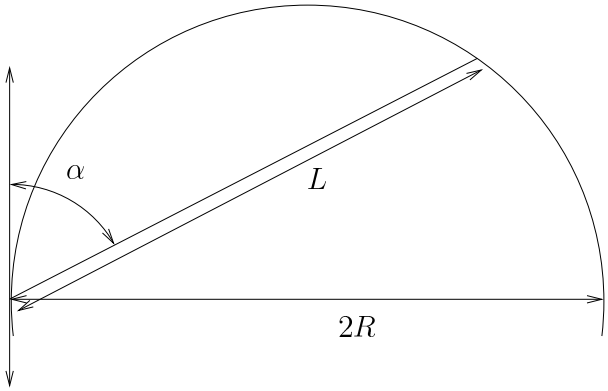


Figure 4.2: Illustration of the flat parts computation on a circle

Proof: A circle of radius R being given, let us consider a chord of length L defining a maximum angle α with the corresponding piece of curve ($0 \leqslant \alpha \leqslant \pi/2$). The values of α and L are related by $L = 2R\sin(\alpha)$. The probability defined earlier is $p(\alpha, L) = \alpha^{R\alpha}$ (expressed as a function of L, it writes down $p(\alpha, L) = \arcsin(L/2R)^{R\arcsin(L/2R)}$). The function $\alpha \mapsto \alpha^{R\alpha}$ shows a minimum for $\alpha = 1/e$. Consequently, $\forall\, \alpha,\ \alpha^{R\alpha} \geqslant e^{-R/e}$.

Thus, if the probability threshold is set to p^*, and if $R < -e\log(p^*)$, then the circles of length R will show no flat part. On the contrary, if $R \geqslant -e\log(p^*)$, the detected flat parts (after the greedy step) in circles of radius R will always show a maximum angle $\alpha = 1/e$ (that is to say 21 degrees, corresponding to an arc of $1/9$ of the total circle), and their length will be $L = 2R\sin(1/e)$. □

Let us remark that p^* only controls the minimum radius under which no flat part will be detected: $-e\log(p^*)$. It appears only through its logarithm and small variations of it will not influence the final result. Now, although for symmetry reasons no piece of circle should be favored by the algorithm, the position of the detected flat parts over a circle strongly depends on the starting point of the discrete curve describing this circle. This makes flat parts of circular curves unreliable in position. As a matter of fact, this will not hinder the recognition of circles, as a a circle matches well with itself, up to any rotation.

4.2 Experiments

4.2.1 Experimental validation of the flat part algorithm

Experimental results are shown in figures 4.4 to 4.9 (original images can be seen on figure 4.3). For each image, the computation time is less than 10 seconds, for a 2GHz standard PC. When images do not show long level lines, the computation time is less than one second.

4.2.2 Flat parts correspond to salient features

Figure 4.10 shows the result of the proposed flat parts detector over all level lines in an image. By "all", we mean that all level lines at all levels with quantization step equal to 1 have been extracted. This permits an exact reconstruction of the original image from the level lines and their corresponding gray levels[130]. Some segments are detected over level lines corresponding to quantization noise (*i.e.* not contrasted level lines over perceptually uniform areas), but these segments actually correspond to small pieces of straight lines. They are not detected any longer when the probability threshold p^* is set to 10^{-10} instead of the standard value (10^{-3}). Flat parts are concentrated along edges. This experiment can be seen as a confirmation that segment lines are actually salient features of images.

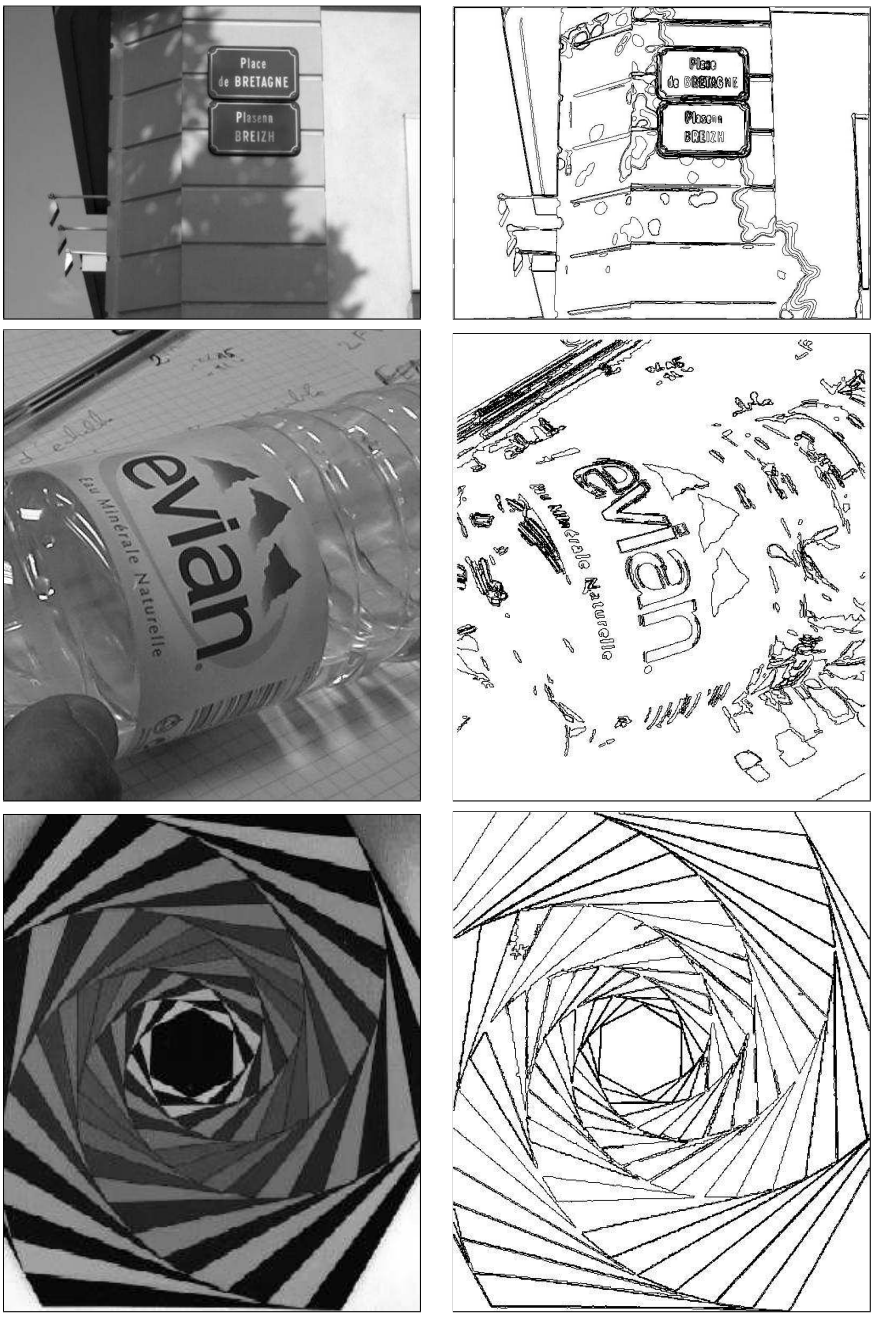


Figure 4.3: Images (left) and meaningful level lines detected with the method described in Chapter 3 (right). Top: bretagne, 413 level lines. Middle: evian, 481 level lines. Bottom: Vasarely, 172 level lines

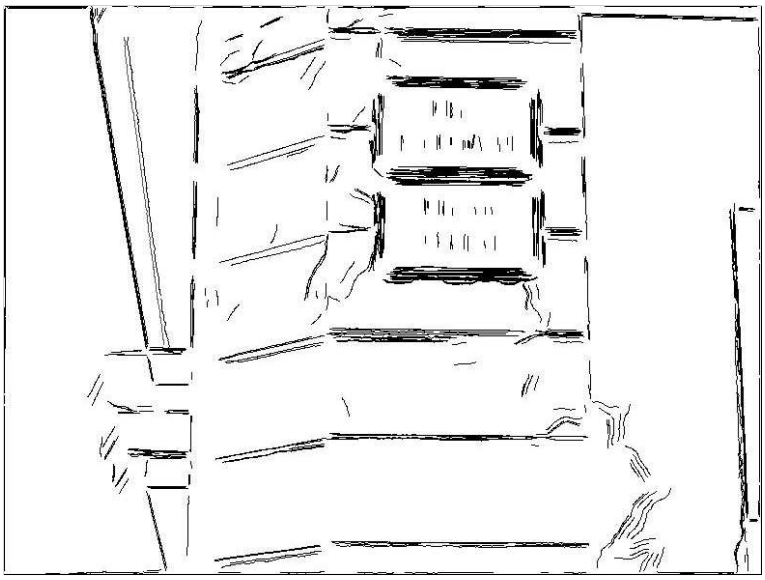


Figure 4.4: Flat parts detection: Bretagne. 1004 detections. Flat parts as small as the ones in the letters of the name of the street are detected (about 10 pixels high). Flat parts in the boundaries of the shadows can be eliminated by dropping the probability threshold, as can be seen on figure 4.5. Nevertheless, these detections actually correspond to small flat parts

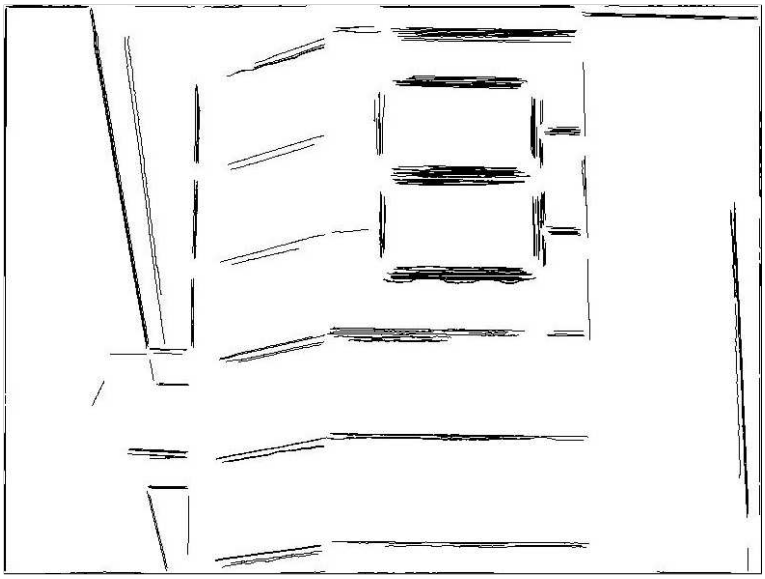


Figure 4.5: Flat parts detection: Bretagne, with $p^* = 10^{-10}$, 417 detections. Letters are too small to be detected, but detected flat parts are very accurate

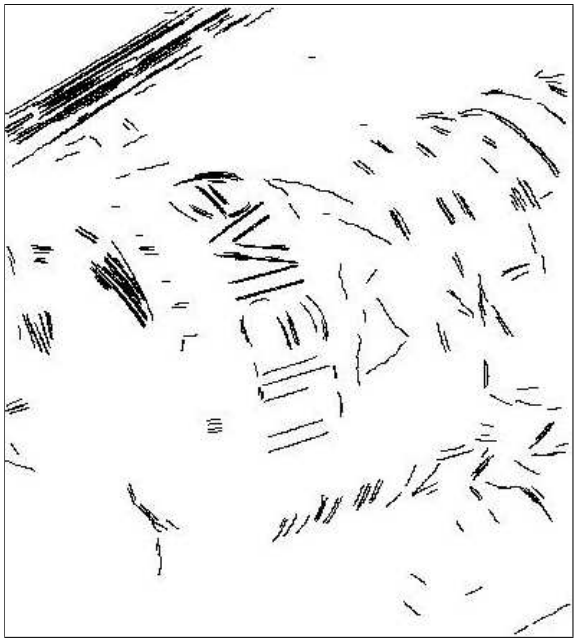


Figure 4.6: Flat parts detection: Evian. 448 detections

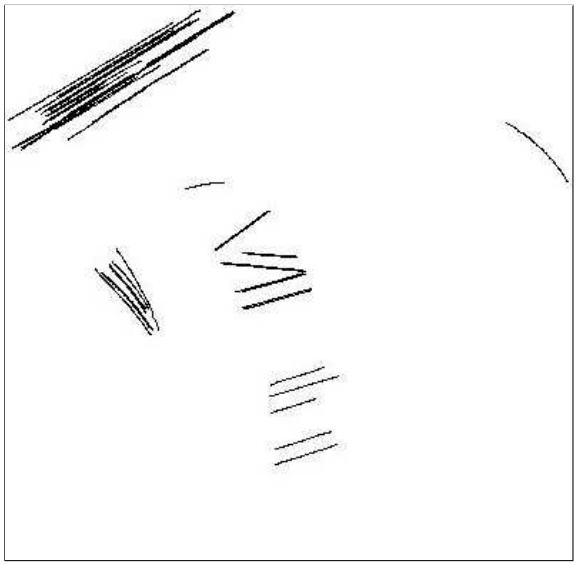


Figure 4.7: Flat parts detection: Evian, with $p^* = 10^{-10}$, 64 detections

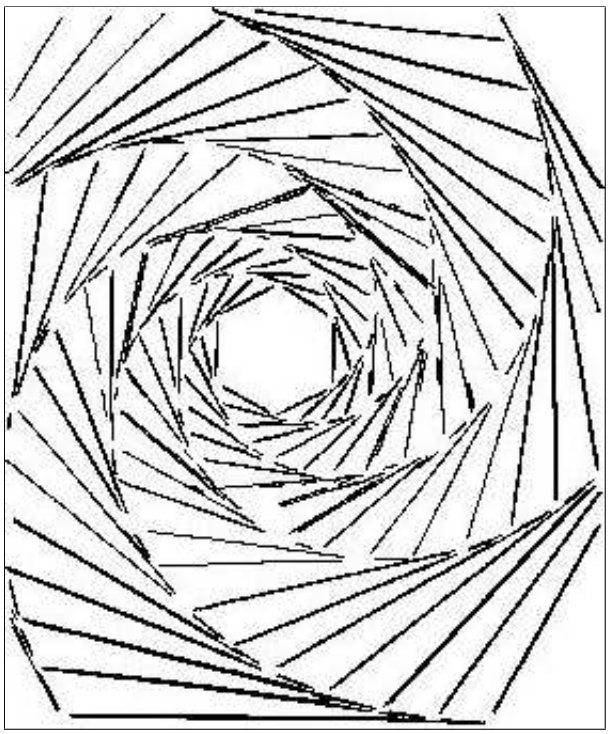


Figure 4.8: Flat parts detection: Vasarely, 774 detections. Each triangle side is correctly detected as a single flat part

Figure 4.9: Flat parts detection: Serena Williams & Puma. Left: Original level lines (425 lines). Middle: $p^* = 10^{-3}$ (675 detections). Right: $p^* = 10^{-10}$ (156 detections). Flat parts on letters are correctly extracted

When comparing to figures 4.4 to 4.7, it can be noticed that, practically, flat parts detection can be restricted to maximal meaningful boundaries. Indeed, the other level lines do not provide valuable information.

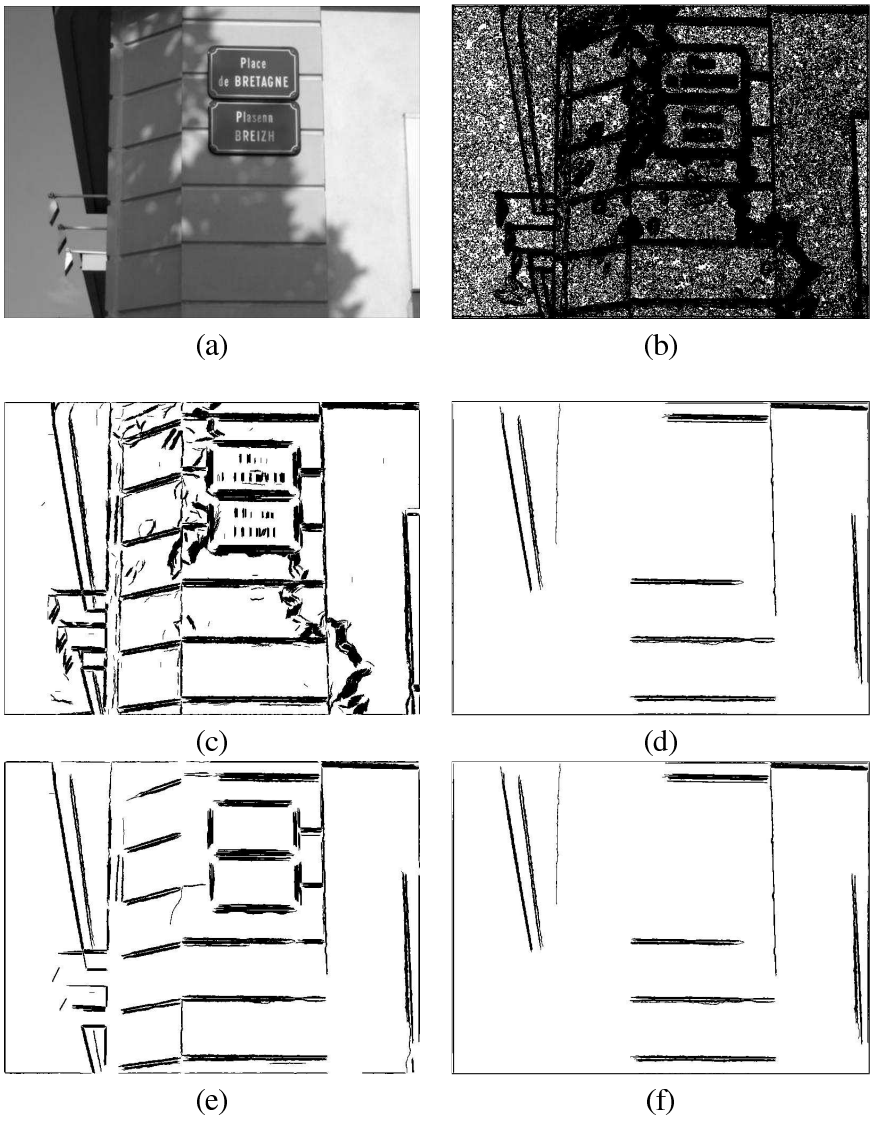


Figure 4.10: Flat parts detection. (a) original image (size: 512×384); (b) 25,755 level lines (quantization step: 1 gray level). They cover the whole image. (c) 20,065 flat parts detected over these level lines (probability threshold p^* has here its standard value: 10^{-3}); (d) flat parts of length larger than 100 pixels among the previous ones; (e) 6,233 flat parts detected over these level lines, when the probability threshold p^* is set to 10^{-10}; (f) flat parts of length larger than 100 pixels among the previous ones. Flat parts appear to be concentrated along edges. These edges appear as thick because a strong gradation of grey can be seen at their location, and thus many parallel pieces of level lines

In his PhD thesis, J.L. Lisani uses flat points in order to build robust semi-local normalisations. Figures 4.12 to 4.15 show a comparison between the proposed flat parts and flat points in the sense of Lisani. See captions for details.

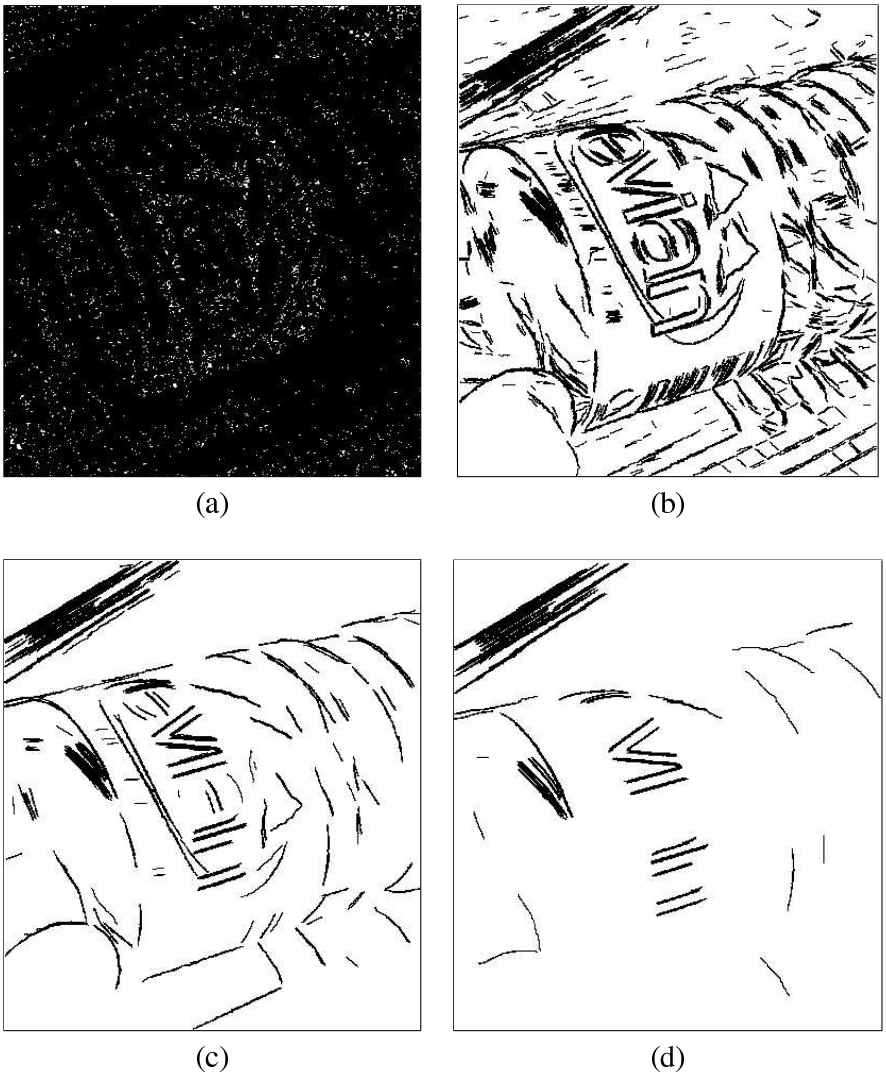


Figure 4.11: Flat parts detection. (a) 90,078 level lines from Evian image (quantization step: 1 gray level); (b) flat parts detections over these level lines (16,533 detections); (c) flat parts detection with $p^* = 10^{-6}$ (4,659 detections); and (d) flat parts detection with $p^* = 10^{-10}$ (2,041 detections). Flat parts are concentrated along edges

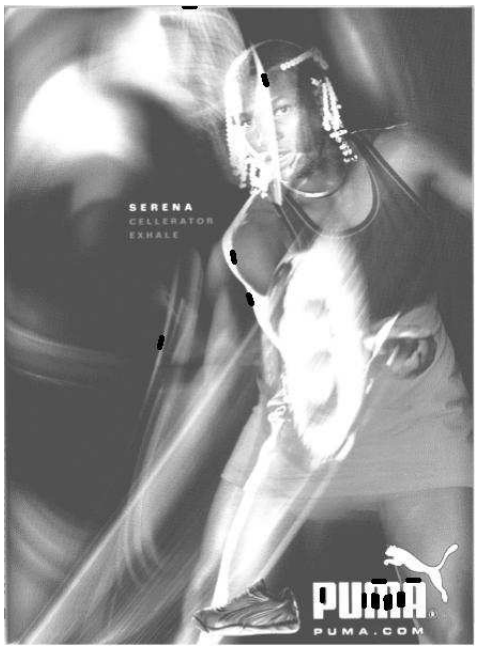


Figure 4.12: J.L. Lisani's flat points: Serena Williams & Puma. 15 flat points are detected. To be compared to the results on figure 4.9

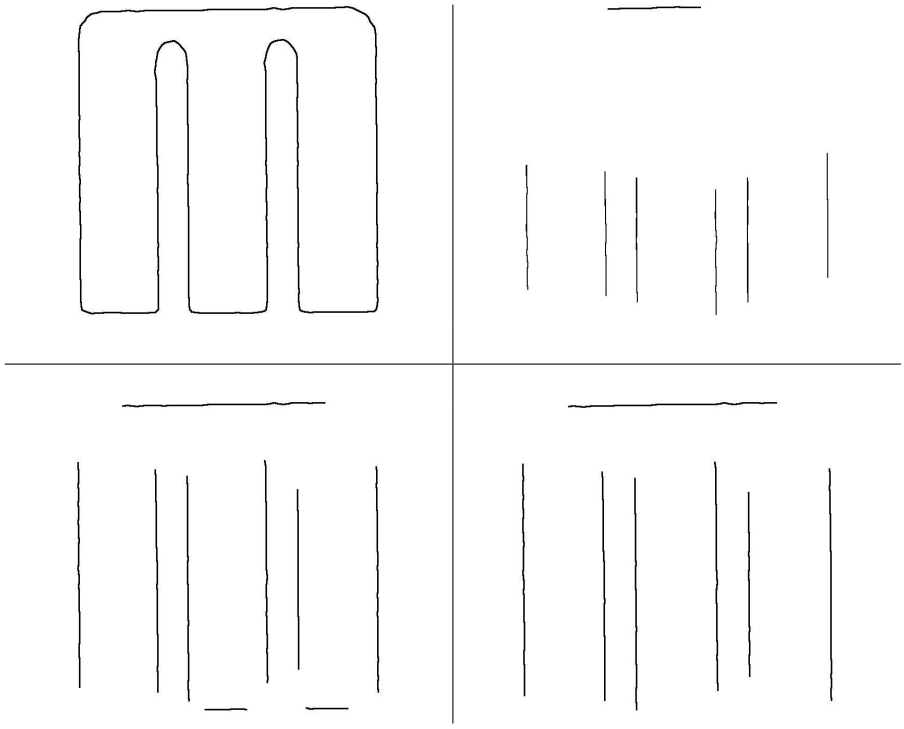


Figure 4.13: Flat points *vs* flat parts: Serena Williams & Puma. From left to right and from top to bottom: considered level line, flat points (7 detections), flat parts with $p^* = 10^{-3}$ (9 detections), flat parts with $p^* = 10^{-10}$ (7 detections). One of the flat parts in the "legs" of the character M is not detected since these curve pieces are too small and pose a sampling problem. Since not *all* chords are tested but a subset of them, endpoints may sometimes be not conveniently distributed

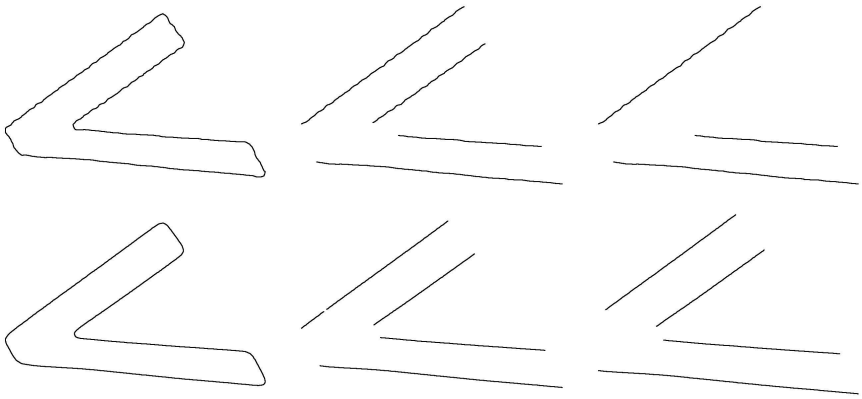


Figure 4.14: Flat points *vs* flat parts: character V in Evian. Top: no smoothing. From left to right: original level line, flat parts with $p^* = 10^{-3}$ (4 detections) and with $p^* = 10^{-10}$ (3 detections). Flat points algorithm does not provide any detection. Bottom: after smoothing. From left to right: original level line, flat parts with $p^* = 10^{-3}$ (5 detections) and flat parts with $p^* = 10^{-10}$ (4 detections). With $p^* = 10^{-3}$, one of the segments is split because of the discretization procedure in the multi-scale test of chords. The Lisani flat points algorithm does not provide any detection

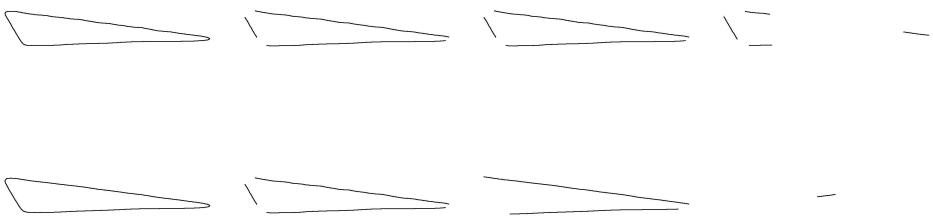


Figure 4.15: Flat points *vs* flat parts: a triangle in Vasarely. Top: no smoothing. From left to right: original level line, flat parts with $p^* = 10^{-3}$ (3 detections) and flat parts with $p^* = 10^{-10}$ (3 detections), and flat points (4 detections). Bottom: after smoothing (see Section 4.3). From left to right: original level line, flat parts with $p^* = 10^{-3}$ (5 detections) and flat parts with $p^* = 10^{-10}$ (2 detections), and flat points (1 detection)

4.3 Curve smoothing and the reduction of the number of bitangent lines

Level lines may be subject to noise, and can have details that are too fine in relation to the essential shape information. Hence, a good shape representation asks for a previous smoothing. Is this smoothing necessary? Quite, from the technological viewpoint, as otherwise there would be too many bitangent lines to level lines and therefore too many geometric codes to a level line. The general framework by which an image or a shape is smoothed at several scales in order to eliminate spurious or textural details and extract its main features is called "Scale Space". The main developments of Scale Space theory in the past ten years involve invariance arguments: indeed, a scale space will be useful for shape recognition only if it is invariant. Let us summarize a series of arguments given in [5]: a scale space computing contrast invariant information must in fact deal directly with the image level lines; in order to be local (not dependent upon occlusions), it must be in fact a partial differential equation (PDE). In order to be a smoothing, this PDE must be of a parabolic kind. Then, the further affine invariance requirement and the invariance with respect to reverse contrast (if a self-dual operator is required, in the mathematical morphology terminology [150]) lead to a single PDE [5]. The PDE characterizing the unique contrast, contrast reversal and special affine invariant scale space is

$$\begin{cases} \frac{\partial u}{\partial t} = |Du|(\text{curv}\, u)^{1/3}, \\ u(x,t) = u_0(x). \end{cases} \tag{4.1}$$

Here $u(t,0) = u_0$ is the initial image, $u(t,x)$ is the image smoothed at scale t and $\text{curv}(u)(x) = \text{div}(\frac{Du}{|Du|})$ denotes the signed curvature of level line passing by x. This equation is equivalent to the "affine curve shortening" [146] of all of the level lines of the image, given by the equation

$$\frac{\partial x}{\partial t} = |\text{Curv}(x)|^{\frac{1}{3}}\mathbf{n}, \tag{4.2}$$

where x denotes a point of a level line, $\text{Curv}(x)$ its curvature and $\mathbf{n}$ the signed normal to the curve, always pointing towards the concavity.
Moisan [122] found a fast algorithm for this curvature motion. For more details on this scheme, one can refer to [122, 97] and to the book [27]. The invariants mentioned mean that the evolution of a shape does not depend upon any affine distortion of the plane. This corresponds to an invariance to all orthographic projections of a planar shape.

Figure 4.16 shows that a slight smoothing by the affine scale space eliminates the sampling effects of a digital image and reduces drastically the number of inflexion points of a shape, without altering its overall aspect. Numerically, the smoothing is slight and stops at the scale $t = 0.5$ at which a circle with radius 0.5 collapses. So the smoothing roughly eliminates details of 1 pixel size.

4.4 Bibliographical notes

4.4.1 Detecting flat parts of curves

In their founding paper [62], M.A. Fischler and R.C. Bowles argue that any curve partitioning technique must satisfy two general principles: stability of the description, and a complete, concise and complexity limited explanation. Smooth sections of curves appear thus to play a major role, because they fit both principles. For instance, Guy and Medioni [75] consider segment lines as *salient* features in images. Considering the detection of flat parts as a multiscale process is not new. It has been used for the more general problem of polygonal approximation of digitized curves (see for example [153]).
Segment or straight line detection is one of the cornerstones of computer vision. Indeed, it is often a preprocessing step of shape recognition, shape tracking [46], vanishing point detection [2], convex shape detection [88], *etc*. Most of the time, straight lines in images are conceived as contiguous edges. Many line detection algorithms therefore require a previous local edge extraction step, such as a Canny's filtering [26]. Hough

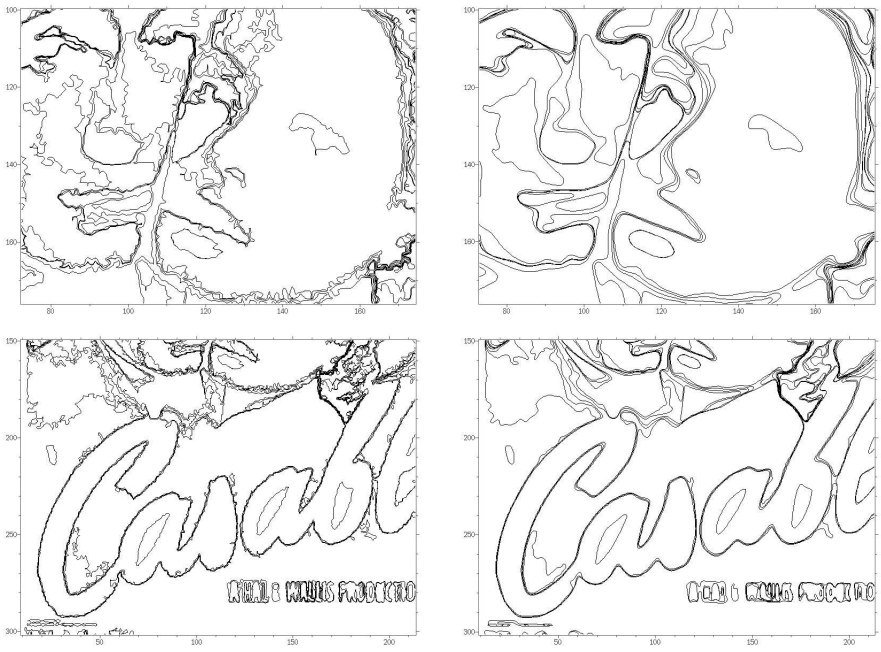


Figure 4.16: Some level lines of a grey level image. Quantization effects and noise can be seen. After a slight smoothing, these effects disappear (on the right)

Transform [81] and algorithms derived from it [87] have been widely studied for that purpose. The goal of these methods is to identify clusters in a particular space (the parameter space of a straight line, either (ρ, θ) with ρ the distance of the line to the origin, and θ the angle between a vector normal to the line and a fixed direction, or (a, b) where a is the slope and b the ordinate of the intersection between the straight line and the ordinate axis). The Hough transform is a voting procedure: every pixel votes for the parameters of the straight line going through it. Another method consists in chaining first the local edges by taking into account connectivity (see for an example [67]), and then in identifying segments among the discrete curves [105]. The main drawbacks of these methods are their number of thresholds (edge detection needs at least a gradient threshold, Hough Transform needs a quantization step for the parameter space discretization and a threshold for the voting procedure) and their computational heaviness and instability (due to local edges chaining). They are moreover very sensitive to noise and to the lack of accuracy of edge detectors: they indeed aim at detecting exact discrete straight lines, in the sense that no outlier edgel is allowed. A concept of fuzzy segments has been proposed [43], but the primary detection is still based on a set of points derived from a local edge detector.

The method presented in this chapter can be viewed as an adaptation to level lines of Desolneux et al. [47], who proposed an a contrario method detecting meaningful alignments in images. A meaningful alignment is a segment where a certain proportion of points have their gradient orthogonal to the same line, up to a given precision. Let us recall the exact definition of a meaningful alignment. A length l segment is ε-meaningful in a $N \times N$ image if it contains at least $k(l)$ points having their direction aligned with the one of the segment, where:

- $k(l)$ is given by: $k(l) = \min\{k \in \mathbb{N}, P(S_l \geqslant k) \leqslant \varepsilon/N^4\}$, and
- $P(S(l) \geqslant k)$ is the probability that, in at least k points in a straight segment of length l, the gradient of the image is orthogonal to the segment, up to a given precision.

The main drawback of this method for segment detection is that it highlights directions and not segments: while the detected straight lines may correspond to the direction of several disjoint segments, gradient direction is allowed to differ between them from the line direction. Estimating the probability that k points among l have

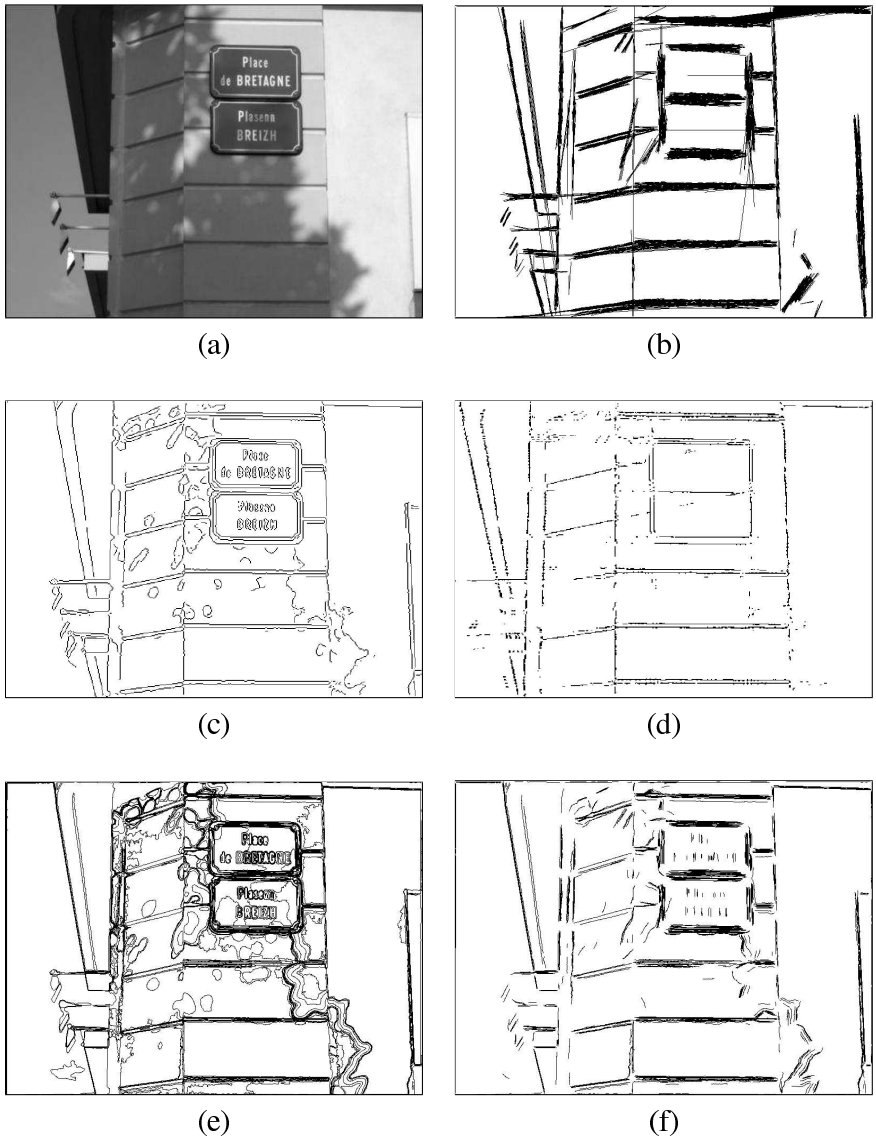
a tangent with the same direction as the chord is not relevant to detect flat parts. In such a model, consecutive alignments are indeed not favored, but are of particular interest in the present setting.

In his PhD thesis [106], Lisani defines "flat points" on curves by using two arbitrary parameters. A "flat point" is the center of a curve segment for which the sum of the angle variations of tangents is small enough (less than 0.2 radian) over a large enough piece of curve (larger than 15 pixels). This algorithm misses many flat points, and does not really detect segments, as several experiments has shown very clearly.

Figure 4.17 shows the results for some of the algorithms which were just discussed. As far as flat parts detection is concerned, Desolneux's alignments are suitable neither for detecting accurate segment directions nor for detecting segment lengths. The naive segment detector based on Hough transform which illustrates the discussion is certainly not the best that can be done using Hough techniques. Nevertheless, even a more clever algorithm would face the same problem as this one: it involves numerous critical parameters (different parameters would drastically change the results). Some isolated points are detected as segments because they fall "by chance" on the same straight line as another more distant segment and therefore collect its votes. Both algorithms (alignments and the Hough transform-based algorithm) are not local enough: that is why segments over the characters in the test image are not detected. Canny's edge detector is well known to suffer from lack of accuracy at edge junctions (where the gradient is badly estimated). Here, this would not be a real issue, since segment lines are searched between junctions, where edges are more accurately detected. Nevertheless, those edge detectors need several critical thresholds.

4.4.2 Scale space and curve smoothing

Since the seminal work of Lamdan *et al.* [99], bitangent lines are well-known to be of high interest to build up such semi-local invariant curve descriptions. The reduction of the number of bitangent lines is linked to curve smoothing, or curve scale space. The modern concept of scale space is due to Witkin [167] and was mainly related to the Gaussian scale space, given by the solution of the heat equation [96]. An interesting shape recognition method using the mean curvature motion was discovered by Mocktharian and Mackworth [127]. The use of curvature-based smoothing for shape analysis is by now well established; founding papers are [10], [127] and [58]. These authors define a multi-scale curvature which is similarity invariant, but not affine invariant. Abbasi *et al.* [1] used the mean curvature motion and an affine length parameterization of the boundary of the solid shapes in order to get an approximately affine shape encoding. Sapiro and Tannenbaum [146] and Alvarez, Guichard, Lions and Morel [5] independently discovered the affine scale space with different approaches. Alvarez et al. proved existence of viscosity solutions to the affine scale space. An existence and regularity theorem was later proved by Angenent, Sapiro and Tannenbaum [7] from which it can be derived that the number of inflexion points decreases under the affine scale space. This result is crucial for shape encoding. Moisan [122] found a fast and fully affine invariant scheme implementing the affine scale space. He also proved the uniform consistency, which, by a result of Barles and Souganidis [16] is a sufficient condition for convergence. The numerical scheme of Moisan was later extended by Cao and Moisan [30] to more general motions by curvature. Very recently the affine erosion scheme was used by Niethammer et al. [133] to compute an affine invariant skeleton of plane curves.



(a) (b)

(c) (d)

(e) (f)

Figure 4.17: Segment detection. (a) original image; (b) maximal meaningful alignments [47]; (c) Canny's edge detector; (d) Points that correspond to an edge and that lie at the same time on a direction detected by voting in the Hough space; (e) local maximal meaningful level lines; (f) result of the proposed algorithm. See text for discussion

Chapter 5

Local and global invariant encoding of shapes

hapters 3 and 4 described the level lines extraction, selection and smoothing procedures, as well as the selection of a few stable, local directions on these curves. These procedures yield shape elements which cannot be directly compared or recognized, since they have undergone an unknown affine transform or a similarity. The classical way to address this problem is *normalization*. We call "affine invariant normalization" a method to build shape representations that are invariant to any planar affine transform $T(x) = Ax + b$, such that $\det(A) > 0$. In other words, an affine invariant normalization transforms a planar shape $\mathcal{F}$ into a normalized shape such that any image of $\mathcal{F}$ by a planar affine transform will lead to the same normalized shape. Notice that shapes related by an axial symmetry are not considered to be equivalent in this framework, and will not yield the same normalized shape. A "similarity invariant normalization" is defined in the same way, where A is just a similarity. Section 5.1 first presents the most classical moment method for affine normalization. It will be experimentally shown that this method is not efficient. In Sect. 5.1.2, a much more accurate normalization method is proposed, involving local and robust features of a level line such as bitangent lines and flat points. This method is applied first to global level lines and then adapted in Sect. 5.2 to pieces of level lines, thus ensuring a robustness of the shape recognition process to occlusions.

5.1 Global normalization and encoding

5.1.1 A global affine invariant normalization method based on moments

Classical shape normalization methods are based on the normalization of its inertia matrix. We shall use Thierry Cohignac's presentation of this method [37]. As will be seen, this method has some drawbacks, that are common to all moment based normalization methods: they rely in the computation of high order moments and are therefore unstable and very sensitive to noise. This will lead us in the next section to propose a global geometric normalization technique based on robust directions (bitangent lines and flat pieces of each level line). So the use of moments based normalization is not recommended. It is, however, simple and elegant and needs to be presented before a more intricate and efficient way is proposed.

Let us denote by $\mathbb{1}_{\mathcal{F}}$ the indicator of a solid shape $\mathcal{F}$. In order to achieve translation invariance of the normalized representation, it may be assumed that $\mathcal{F}$ has been previously translated such that its barycenter is in the origin of the image plane. Hence, the moment of order (p, q) (p and q natural integers) of $\mathcal{F}$ is defined by

$$\mu_{p,q}(\mathcal{F}) = \int_{\mathbb{R}^2} x^p y^q \mathbb{1}_{\mathcal{F}}(x, y) dx dy.$$

Let $S_{\mathcal{F}}$ be the following 2×2 positive-definite, symmetric matrix

$$S_{\mathcal{F}} = \frac{1}{\mu_{0,0}} \begin{pmatrix} \mu_{2,0} & \mu_{1,1} \\ \mu_{1,1} & \mu_{0,2} \end{pmatrix},$$

where $\mu_{i,j} = \mu_{i,j}(\mathcal{F})$. By the uniqueness of Cholesky factorization [68], $S_\mathcal{F}$ may be uniquely decomposed as $S_\mathcal{F} = B_\mathcal{F} B_\mathcal{F}^{\mathrm{T}}$ where $B_\mathcal{F}$ is a lower-triangular real matrix with positive diagonal entries.

DEFINITION 5.1 *The* pre-normalized shape *associated to $\mathcal{F}$ is the shape $\mathcal{F}' = B_\mathcal{F}^{-1}(\mathcal{F})$.*

The aim is to prove that the pre-normalized solid shape is invariant to affine transforms, up to a rotation.

LEMMA 5.1 *Let A be a non-singular 2×2 matrix. Then $S_{A\mathcal{F}} = A S_\mathcal{F} A^{\mathrm{T}}$.*

Proof: Let a, b, c and d be real numbers such that:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

The moment of order $(2,0)$ associated to the solid shape $A\mathcal{F}$ is

$$\begin{aligned} \mu_{2,0}(A\mathcal{F}) &= \det(A) \int_{\mathbb{R}^2} (ax+by)^2 \mathbb{1}_\mathcal{F}(x,y) dxdy \\ &= \det(A)(a^2\mu_{2,0} + 2ab\mu_{1,1} + b^2\mu_{0,2}). \end{aligned}$$

The same computation for moments of order $(0,2)$ and $(1,1)$ yields

$$\begin{aligned} \mu_{0,2}(A\mathcal{F}) &= \det(A)(c^2\mu_{2,0} + 2cd\mu_{1,1} + d^2\mu_{0,2}), \\ \mu_{1,1}(A\mathcal{F}) &= \det(A)(ac\mu_{2,0} + bd\mu_{0,2} + (ad+bc)\mu_{1,1}). \end{aligned}$$

Since $\mu_{0,0}(A\mathcal{F}) = \det(A)\mu_{0,0}$, one can easily check that $S_{A\mathcal{F}} = A S_\mathcal{F} A^{\mathrm{T}}$. □

LEMMA 5.2 *Let X_0 be a 2×2 invertible matrix. Then, for any 2×2 matrix X: $XX^{\mathrm{T}} = X_0 X_0{}^{\mathrm{T}}$ if and only if there exists an orthogonal matrix Q such that $X = X_0 Q$.*

Proof: Since X_0 is invertible, $XX^{\mathrm{T}} = X_0 X_0{}^{\mathrm{T}}$ if and only if $X_0^{-1} X (X_0^{-1} X)^{\mathrm{T}} = \mathrm{Id}_2$. Letting $Q = X_0^{-1} X$ yields the result. □

PROPOSITION 5.1 *The pre-normalized solid shape is invariant to any invertible, planar, linear transform $(x,y)^{\mathrm{T}} \mapsto A(x,y)^{\mathrm{T}}$, up to an orthogonal transform. Moreover, if $\det(A) > 0$, the invariance holds up to a rotation.*

Proof: Since A is a 2×2 non singular matrix, following Lemma 5.1, $S_{A\mathcal{F}} = A S_\mathcal{F} A^{\mathrm{T}}$. By letting $B_\mathcal{F}$ be the lower-triangular matrix of Cholesky's decomposition of $\mathcal{F}$, it follows that $S_{A\mathcal{F}} = AB_\mathcal{F}(AB_\mathcal{F})^{\mathrm{T}}$. Now, since $S_{A\mathcal{F}}$ is a 2×2 positive-definite, symmetric matrix, Cholesky factorization yields $S_{A\mathcal{F}} = B_{A\mathcal{F}} B_{A\mathcal{F}}{}^{\mathrm{T}}$, where $B_{A\mathcal{F}}$ is a 2×2 non-singular, lower-triangular real matrix. Then, by Lemma 5.2, $B_{A\mathcal{F}} = AB_\mathcal{F} Q$, where Q is a 2×2 orthogonal matrix. Hence, $B_{A\mathcal{F}}^{-1} A\mathcal{F} = (AB_\mathcal{F} Q)^{-1} A\mathcal{F} = Q^{-1} B_\mathcal{F}^{-1} A^{-1} A\mathcal{F} = Q^{-1} B_\mathcal{F}^{-1} \mathcal{F}$, what proves the invariance of $\mathcal{F}' = B_\mathcal{F}^{-1}\mathcal{F}$ to planar isomorphisms, up to an orthogonal transform. Finally, notice that if $\det(A) > 0$, then $\det(Q) > 0$. □

A closed form for $B_\mathcal{F}^{-1}$ in terms of the moments of $\mathcal{F}$ can be computed by taking the inverse of $B_\mathcal{F}$, the lower-triangular matrix given by the Cholesky decomposition of $S_\mathcal{F}$,

$$B_\mathcal{F}^{-1} = \sqrt{\mu_{0,0}} \begin{pmatrix} \frac{1}{\sqrt{\mu_{2,0}}} & 0 \\ -\frac{\mu_{1,1}}{\mu_{2,0}\sqrt{\mu_{0,2} - \frac{\mu_{1,1}^2}{\mu_{2,0}}}} & \frac{1}{\sqrt{\mu_{0,2} - \frac{\mu_{1,1}^2}{\mu_{2,0}}}} \end{pmatrix}.$$

The pre-normalized solid shape $\mathcal{F}' = B_{\mathcal{F}}^{-1}\mathcal{F}$ is then an affine invariant representation of $\mathcal{F}$ *modulo* a rotation. In order to obtain a full affine invariant representation, only a reference angle is needed. This can be achieved, for instance, by computing

$$\varphi = \mathrm{Arg}\left(\int_0^{2\pi}\int_0^{+\infty} \mathbb{1}_{\mathcal{F}'}(r,\theta)e^{i\theta}rdrd\theta\right),$$

then rotating $\mathcal{F}'$ by $-\varphi$. Notice that this rotation normalization method fails when $\mathcal{F}'$ exhibits a central symmetry. However, unlike a classical rotation normalization computing the direction of the principal axis, it has the advantage of assigning the same weight to all points in $\mathcal{F}'$, and hence to be more robust to the noise affecting its boundary.

Putting all the steps together, the affine invariant normalization of a solid shape $\mathcal{F}$ is the set of points (x_N, y_N) given by

$$\begin{pmatrix} x_N \\ y_N \end{pmatrix} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} B_{\mathcal{F}}^{-1} \begin{pmatrix} x-\mu_{1,0} \\ y-\mu_{0,1} \end{pmatrix},$$

for all $(x,y) \in \mathcal{F}$.

As can be seen in Figure 5.1, a classical problem of this kind of normalization is its lack of robustness. Too strong deformations lead to a bad estimation of the moments.

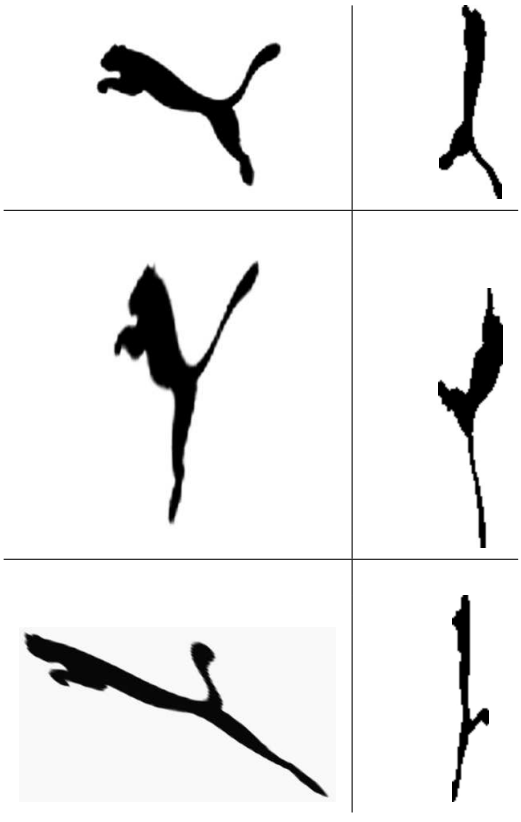


Figure 5.1: T. Cohignac's normalization. Original images (on the left) and affine normalization using the moments (on the right). The middle and bottom original image were mapped from the top original image up to an affine transform. Even in this ideal framework, the obtained normalized solid shapes are not superposable at all: the moment normalization is not robust. Compare with the local normalization proposed in the next section (the middle original image was mapped up to the same transform as on Figure 5.3)

5.1.2 Geometric global normalization methods

The geometric global normalization method described thereafter is based on robust directions given by the bitangent lines and the flat pieces of a solid shape boundary or level line $\mathcal{L}$. In the previous method, the second order moments of the moment based global normalization were used to find principal directions of the shape. The bitangent lines and flat parts will now play that role and lead to a much more reliable geometric normalization. A similarity invariant and an affine invariant global normalization methods are proposed here. The best way to describe such methods is to directly give their algorithm, which is self-explanatory. In the following we consider direct Euclidean parameterization for level lines, as usual.

Similarity invariant normalization

For each shape $\mathcal{F}$ with boundary $\mathcal{L}$, and for all "robust" straight line $\mathcal{D}$ computed from $\mathcal{L}$:

1. Translate $\mathcal{F}$ so that its barycenter becomes the origin of the plane.
2. Scale $\mathcal{F}$ so that its boundary has unit length.
3. Rotate $\mathcal{F}$ with respect to the origin so that the robust direction is horizontal.
4. Define the starting point of the parameterization of $\mathcal{L}$ as the intersection with positive ordinate between the vertical axis and the boundary of the solid shape. In case of ambiguity, choose the closest one to the origin.

Affine invariant normalization (positive determinant)

The procedure is illustrated in Figure 5.2. For each "robust" straight line $\mathcal{D}$ computed from $\mathcal{L}$:

1. Consider the straight line passing through the barycenter G of $\mathcal{F}$, which is parallel to $\mathcal{D}$. Consider the intersection between $\mathcal{F}$ and the half-plane defined by this straight line which does not contain $\mathcal{D}$; call G_1 its barycenter, and G_3 the barycenter of the complementary part of $\mathcal{F}$.
2. Now consider the straight line passing through G_1 and G_3. It splits the solid shape into two parts, let G_2 and G_4 be their barycenter, such that $(\overrightarrow{G_3G_1}, \overrightarrow{G_2G_4})$ is directly oriented. (The lines G_1G_3 and G_2G_4 intersect at G.)
3. Points $\{G, G_1, G_2\}$ define an affine basis. Normalize $\mathcal{F}$ by applying to it the affine transform mapping $\{G, G_2, G_1\}$ into $\{(0,0), (1,0), (0,1)\}$.
4. Define the starting point of the parameterization of $\mathcal{L}$ as the intersection with positive ordinate between the vertical axis and the boundary of the normalized solid shape. In case of ambiguity, choose the closest one to the origin.

The proof of the next proposition is straightforward from the preceding algorithms.

PROPOSITION 5.2 *Let $\mathcal{L}_1$ and $\mathcal{L}_2 = A\mathcal{L}_1$ be two curves such that $\mathcal{L}_1$ is deduced from $\mathcal{L}_2$ by a similarity (resp. affine) transform, denoted by A. Then the sets of all normalized geometric curves obtained by the above normalization algorithms, applied to all bitangent lines, are identical.*

Proof: There is by A a one-to-one correspondence between the bitangent lines of $\mathcal{L}_1$ and $\mathcal{L}_2$ and the two above algorithms then describe a similarity (resp. affine) invariant procedure, leading to identical normalized shapes. □

The former result was enounced for bitangent lines only, as the robust lines also obtainable from flat pieces of the curves are not *stricto sensu* similarity or affine invariant. Notice, however, that the use of flat zones is

unavoidable to encode convex shapes, which have no bitangent lines. Moreover, under reasonable zoom factors, flat parts are preserved. Flat parts are often detected as tangent lines at inflexion points (which are conserved by affine transforms).

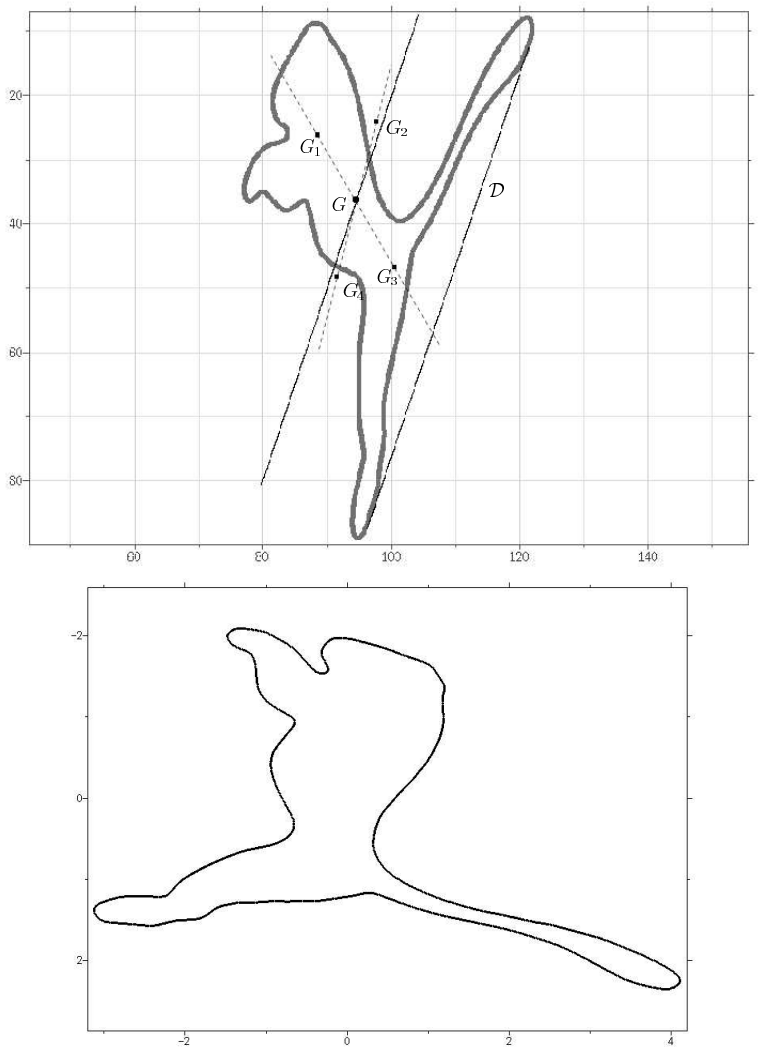


Figure 5.2: Global affine invariant normalization based on the bitangent line $\mathcal{D}$. Top: definition of points G_1, G_2, G_3 and G_4. Bottom: the normalized solid shape

Figure 5.3 shows an example of global affine invariant normalization. The considered shapes are the same as in Figure 5.1. Notice that the normalization is much more stable than in the moment based approach.

5.2 Semi-local normalization and encoding

The necessity of a local shape encoding was emphasized enough in this book. So the preceding sections about global encoding are mere essays towards a local one. This will be actually a simple adaptation and it will be shorter and clearer to directly proceed and describe the algorithms.

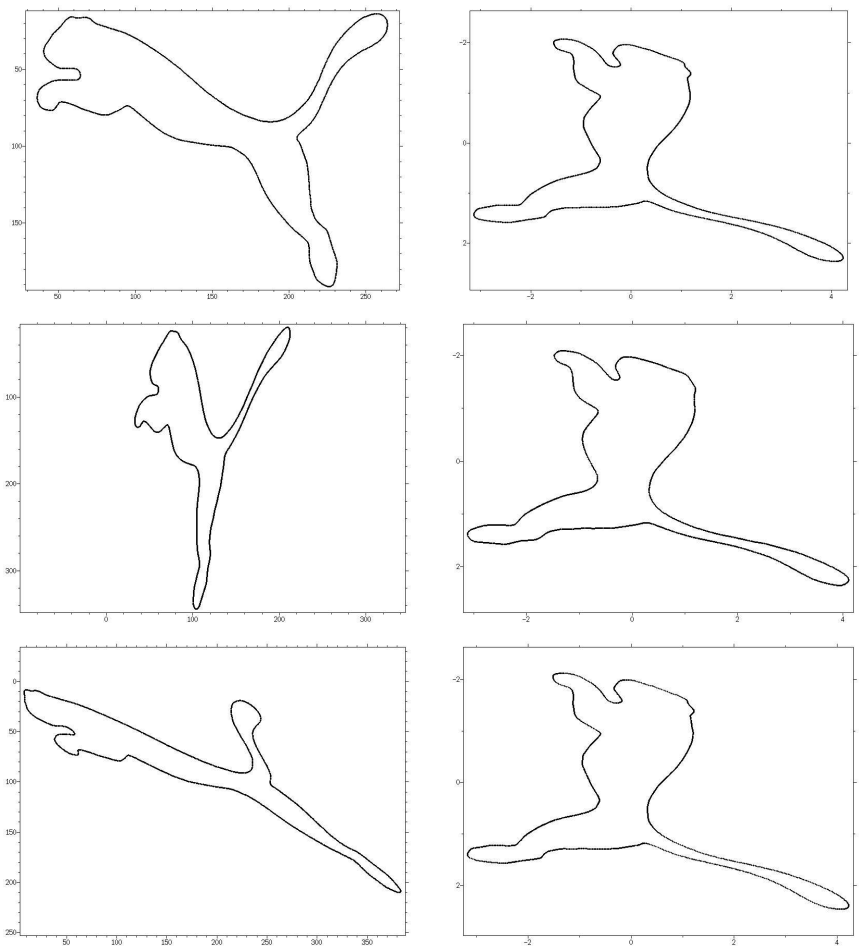


Figure 5.3: Global affine invariant normalisation based on robust directions. Left column: Boundaries of the original image (on top) and of two affine deformations of it (the same ones as in Figure 5.1). Right column: corresponding affine normalizations based on a bitangent line. The normalized shapes are very close; this is not the case with the invariant moment method

Similarity invariant normalization and encoding

Given a level line $\mathcal{L}$, for each flat piece, or for each bitangent line, do the following (this procedure is illustrated in Figure 5.4):

a) Call P_1 the first tangency point and P_2 the other one (for flat pieces, P_1 and P_2 are the endpoints of the detected flat segment). Consider the tangent line $\mathcal{D}$ containing these points;

b) Call $\mathcal{P}_1$ the first tangent line to $\mathcal{L}$ which is orthogonal to $\mathcal{D}$, starting from P_1 in the negative direction. Call $\mathcal{P}_2$ the first tangent line to $\mathcal{L}$ which is orthogonal to $\mathcal{D}$, starting from P_2 in the positive direction.

c) Find the intersection points between $\mathcal{P}_1$ and $\mathcal{D}$, and between $\mathcal{P}_2$ and $\mathcal{D}$. Call them R_1 and R_2, respectively;

d) Store the *normalized* coordinates of N equi-distributed points over an arc on $\mathcal{L}$ of length $F \cdot \|R_1R_2\|$, centered at C, the intersection point of $\mathcal{L}$ with the perpendicular bisector of $[R_1R_2]$ (the first intersection starting from P_1). By "normalized coordinates", one has to understand coordinates in the similarity invariant frame defined by points R_1, R_2 mapped to $(-\frac{1}{2}, 0), (\frac{1}{2}, 0)$, respectively.

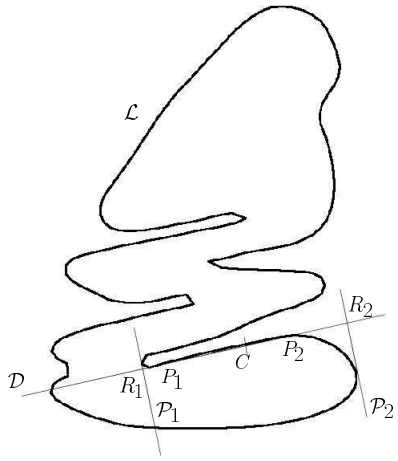


Figure 5.4: Similarity invariant semi-local encoding based on a flat part of straight line $\mathcal{D}$

Two implementation parameters, F and N, are involved in this normalization procedure. The value of F determines the normalized length of the shape elements. It has to be chosen having in mind the following trade-off: If F is too large, shape elements will be too long to deal with occlusions, while if it is too small, shape elements will not be discriminatory enough. The choice of F faces then a classical dilemma in shape analysis, addressed in the bibliographical notes of this chapter (Sect. 5.3): locality *versus* globality of shape representations. The choice of N is less critical from the shape representation viewpoint, since it is just a sampling precision parameter. Its value is to be chosen as a compromise between accuracy of the shape element representation, and computational load.

Figure 5.5 shows some shape elements extracted from a single boundary, taking $F = 5$ and $N = 45$. Notice that the representation is quite redundant, and yields shape elements describing the boundary over a wide range of scales. This redundancy increases the possibility of recognition when shapes are degraded or subject to partial occlusions.

All the experiments in Chapter 7 concerning matching based on this semi-local encoding (section 7.1) were carried out using $F = 5$ and $N = 45$. These parameters can be fixed once for all, and they are not to be tuned by the user.

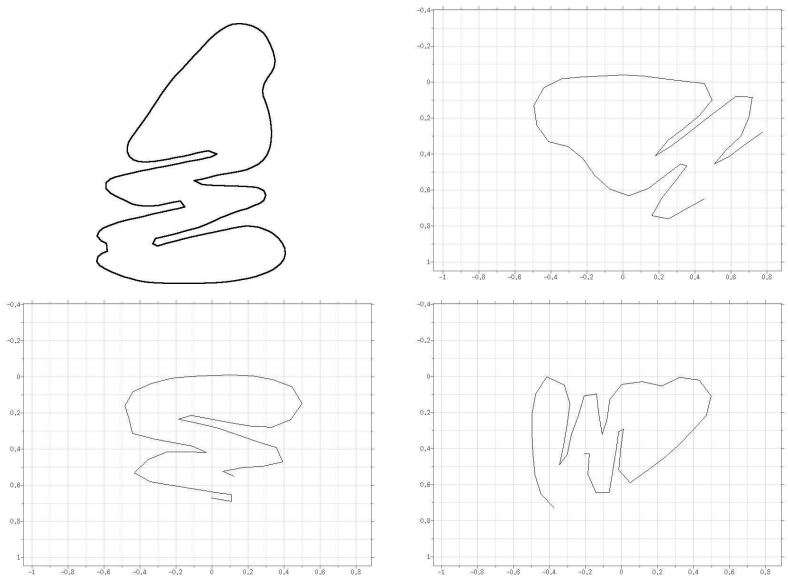


Figure 5.5: Example of semi-local similarity invariant encoding. The line on the left generates 19 shape elements ($F = 5$, $N = 45$). Twelve of them are based on bitangent lines, the other ones are based on flat pieces. The representation is of course redundant. Here are displayed three normalized shape elements, two deriving from bitangent lines, and one from a flat piece

Affine invariant normalization/encoding

The affine invariant representation of a level line $\mathcal{L}$ is computed by applying the following procedure for each flat piece or bitangent of $\mathcal{L}$ (this procedure is illustrated in Figure 5.6):

a) Call P_1 the first tangency point and P_2 the other one (for flat pieces, P_1 and P_2 are the endpoints of the detected flat segment). Consider the tangent line $\mathcal{D}$ to these point;

b) Starting from P_2, find the next tangent to $\mathcal{L}$ which is parallel to $\mathcal{D}$. Call it $\mathcal{D}'$;

c) Consider the straight lines which are parallel to $\mathcal{D}$ and lay at $1/3$ and $2/3$ of distance from $\mathcal{D}$ to $\mathcal{D}'$. Call them $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively;

d) Starting from P_2, find the next intersection points between $\mathcal{L}$ and $\mathcal{D}_1$, and $\mathcal{L}$ and $\mathcal{D}_2$. Consider the straight line $\mathcal{T}_1$ defined by these two points.

e) Starting from P_1, find the previous tangent to $\mathcal{L}$ parallel to $\mathcal{T}_1$, and call it $\mathcal{T}_2$;

f) Define points R_1, R_2, and R_3 as the intersections between $\mathcal{D}$ and $\mathcal{T}_2$, $\mathcal{D}$ and $\mathcal{T}_1$, and $\mathcal{D}'$ and $\mathcal{T}_2$, respectively;

g) Points R_1, R_2, R_3 define an affine basis . The affine normalization is fixed by mapping $\{R_1, R_2, R_3\}$ into $\{(0,0),(1,0),(0,1)\}$ if $\{R_1, R_2, R_3\}$ is a direct frame, and into $\{(0,0),(1,0),(0,-1)\}$ if not.

h) Encoding: consider the intersection point between $\mathcal{L}$ and the straight line equidistant from $\mathcal{D}$ and $\mathcal{D}'$ (the first one starting from P_2). Call it C. Normalize the portion of $\mathcal{L}$ having normalized length $F/2$ at both sides of C. Store N equi-distributed points over the normalized piece of curve.

As for the similarity invariant normalization, implementation parameters were fixed once for all to $F = 5$ and $N = 45$. Figure 5.7 shows all shape elements extracted from a single boundary for this choice of parameters. Notice that the encoding is less redundant than for the similarity encoding procedure. This is due to the fact that the construction of affine invariant local frames imposes more constraints on the curve than the one for similarity invariant frames.

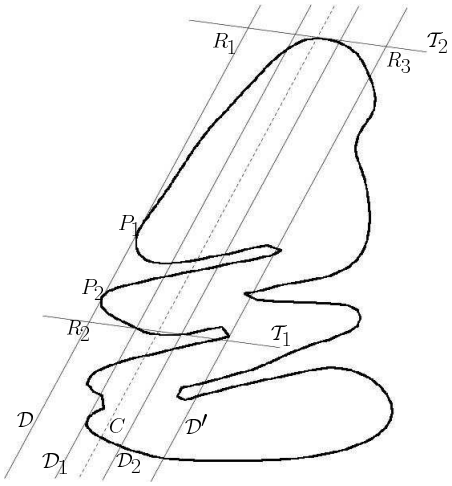


Figure 5.6: Affine invariant semi-local encoding. The encoded shape element is based on the tangent to the flat piece between marks

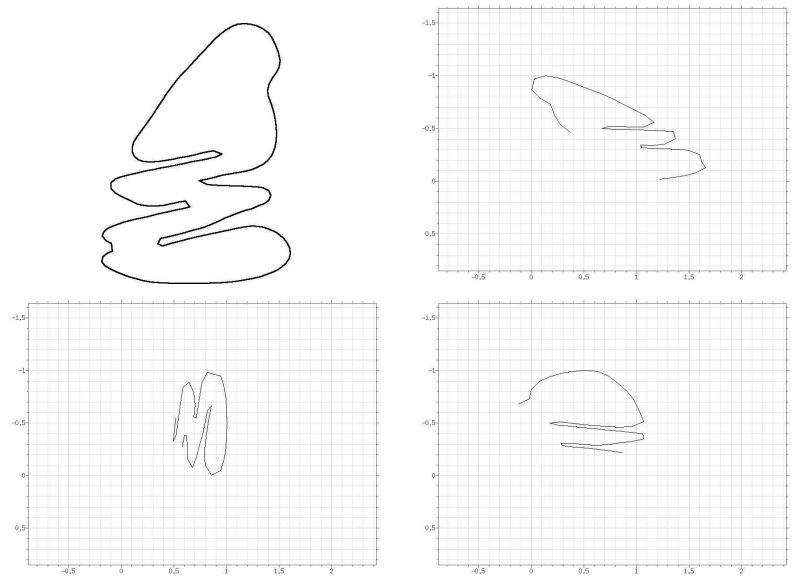


Figure 5.7: Example of semi-local affine invariant encoding. The line on the left generates 7 shape elements ($F = 5$, $N = 45$); three of them are represented here

Typical number of shape elements in images

The number of shape elements of a grey level image depends on the complexity of its level lines. Indeed, the number of shape elements is roughly proportional to the number of inflexion points. Textured images have in general many shape elements since their level lines are quite complex. To give an order of magnitude, the level lines of a database of 23 natural images of different type were encoded using the similarity encoding procedure described above. The level lines were respectively:

1. all meaningful boundaries,
2. only maximal meaningful boundaries,
3. maximal meaningful boundaries with local contrast,
4. cleaned (see Sect. 3.2.3) maximal meaningful boundaries with local contrast.

The number of shape elements per pixel and the CPU time of the encoding per pixel were measured, see Table 5.1.

Table 5.1: Number of shape elements encoded by the similarity semi-local encoding algorithm. Using local boundaries and the cleaning procedure makes the encoding much faster. In addition, the shape elements dictionary are shorter, but they experimentally contain all characteristic pieces of objects boundaries. The matching phase complexity is directly proportional to the number of pairs of shape elements, one taken in each image to be match. Thus, this simplification is algorithmically quite rewarding.

	#shape elements/pixel	CPU (s)/pixel
all MB	0.1458	0.0024
maximal MB	0.0528	0.0006
local MB	0.0310	0.0004
cleaned local MB	0.0132	0.0002

The typical time for encoding the meaningful boundaries is between 10s and 1min. The gain from meaningful boundaries to maximal meaningful boundaries is obvious and due to the elimination of redundancies in the level lines tree. The gain by involving local meaningful boundaries is experimental since it is possible to construct images with more local meaningful boundaries than maximal meaningful boundaries. Since the cleaning procedure removes some parts of the level lines, the encoding is logically faster and the shape elements dictionary shorter.

5.3 Bibliographical notes

The level lines extraction/smoothing/geometric encoding as described in this book was first introduced by Lisani *et al.* [106, 107]; the third geometric encoding stage described in this chapter is inspired from this reference, Rothwell's work on invariant indexing [142]. The aim of the next subsections is to review and attempt to classify a wide number of antecedent shape encoding methods. When adequate, usual drawbacks will be pointed out, without entering into details: sensitivity to noise, to occlusion, or to deformation. This will hopefully make more understandable the choice of the single geometric method presented in detail here.

5.3.1 Geometric invariance and shape recognition

Let us first review some references about the invariance issue in shape recognition. When shapes are subject to weak perspective distortions, the human perception is still able to recognize them. The geometric invariance requirement for shape recognition was already discussed in chapter 1, section 2.1.1. We claimed that in a general setting, affine invariance should be considered, while similarity invariance could be enough for a large class of particular applications. Such a claim was based on the following arguments and articles:

- Projective transforms are shown not to behave well with regard to shape matching, because they permit to map a large class of curves arbitrarily close to a circle, and thus to map a finite number of curves arbitrarily close to a given curve (for example, a rabbit and a duck are "almost" projective equivalent [11, 12]).

- Despite some interesting attempts [60], there is no practical way to define a projective invariant local smoothing. From this viewpoint, affine invariant smoothing is the "best" possible [5].

- Since projective transforms are differentiable, they can be locally approximated by affine transforms (for which invariant smoothing is well defined), and these approximations are particularly accurate under weak perspective distortion.

So affine distortions have to be thought of as local distortions. This is not really a restriction, since the locality of shape representation was already required in order to deal with occlusions and with the figure-background problem (see also Chapt. 1, Sect. 2.1.1).

5.3.2 Global features and global normalization

The simplest recognition methods are global, in the sense that the extracted features are computed over the whole solid shape. Since they mix global and local information, they are sensitive to occlusions (a part of the solid shape is hidden) or insertion (a part is added to the solid shape). This makes them inappropriate for general applications, and restrict their use to a few specific applications, where the observed objects do not overlap. The global features are in general scalar numbers computed over the whole solid shape. In the case of closed curves, Fourier descriptors [98, 102, 140, 171] or invariant moments [57, 128] (following Hu [82]) can be used. Affine invariant scalars for global shape representation can also be derived from wavelet coefficients [94, 151]. Using wavelets allows one to capture some local information of the shape, but not to the point of being able to deal with occlusion (the invariant scalars are computed by using coefficients from different scales). Another well-known, moment related, global method is the *modal matching*, by Sclaroff and Pentland [149]. In this method, a physical elastic model of the solid shapes is considered. Shapes are represented by the ordered set of their eigenvalues associated to the elastic model. This method permits relatively realistic shape deformations where the thin parts of the shape can alter more than the bulk.

An original approach using size functions was proposed by Frosini *et al.* [63, 64]. Size functions can be seen as tools to get information about the topology of any graph. Applied to shape recognition, the size function theory leads to nearly-invariant descriptors, which can be well adapted to perceptual matching since they rely on structural information. Methods based on moments or Fourier descriptors, as well as size function methods, face the same problem: how to define the relative weights of each moment, or size function, in a shape comparison distance? This choice is in general arbitrary, or based on *ad hoc* arguments. Robustness against noise is another aspect of this problem. Since high order moments (or high frequency *modes* for the modal matching method) represent details or fine information of the shape, they can be contaminated by noise and they should not be considered. But up to what order should moments be considered?

Moments based normalization methods, like the one presented in Section 5.1.1, have been extensively used in shape recognition. As noticed before, these methods suffer from two stability problems. First, because of the dependence on second order moments, the points in the contour of the shape strongly influence the result, making the normalization quite sensitive to contour deformations. This effect can be reduced by considering robust norms like Geman-McClure's ρ-function [66] for the estimation of the principal axis, instead of the standard quadratic norm [42, 39]. Second, an error in the identification of the principal axis when the shape eigenvalues are close may yield completely different normalizations.

More stable global normalization methods can be built by considering bitangents lines, like in the geometric normalization method proposed in Section 5.1.2. In [134], affine invariant frames for global shape normalization are built, by considering the pair of tangency points of the curve with the bitangent line, and an extra point

which can be the barycenter of the solid shape. This should not be as stable as the geometric global normalization proposed here, since the the position of the bitangency points is not as robust as the direction of the bitangent line.

The scale-space representation of level lines can also be used to derive invariant representations. One such method can be found in Alvarez *et al.* [6], where shape invariants are based on the evolution of area and perimeter of the solid shapes surrounded by the level line undergoing the affine scale space. Let us describe the seminal work by Mokhtarian and Mackworth [127]. A shape (thought of as a Jordan curve) is smoothed by curvature motion. At each scale, the smoothed curve is reparameterized by the normalized arc length, and the position of inflexion points (zero-crossings of the curvature) is tracked. If σ denotes the scale and s the corresponding normalized arc length, the proposed multiscale representation of the shape consists in the set of 2-tuples (s_i, σ_i), corresponding to the position and the scale at which two inflexion points meet and vanish. The corresponding binary image in the (s, σ) plan has been called the *Curvature Scale Space* and is a similarity invariant representation. It can also be robust to noise, if one only considers the information given by the scale space for scales larger than an *ad hoc* or arbitrarily fixed threshold. At first sight, this method seems to be able to deal with occlusion, since curvature is a local property of curves. This is not the case, however, since at each scale curves are reparameterized by the normalized arc length, and occlusions or insertions can drastically modify the positions of points (s_i, σ_i).

5.3.3 Local and semi-local features

While global features are in general defined to be geometrically invariant up to, at least, rigid transforms, the local or semi-local features defined in the shape recognition literature can be invariant or not.
Commonly used non invariant features are, for instance, sets of edges [114, 115]. Groups of features are more informative than individual local features, and consequently enhance the matching stages: chained edges [169] or edgels [135] (an edge element with a direction) can be considered.
In order to achieve (geometrical) invariant recognition, non invariant features must be compared by means of strategies dealing with invariance, thus leading to time consuming algorithms. Non invariant features will not be further discussed.
Invariant local features may be computed directly on the image, or after the shape has been extracted. Features can be differential or integro-differential invariants at some special points (like corners [148]) or regions (*e.g.* coherent regions [19, 166]) of the image. The computation of differential invariants is very unstable, even after smoothing the image, since it involves high order derivatives.
Weiss [164] proposes local projective invariants requiring the computation of fourth order derivatives of the curves. This is of course out of range for contours of solid shapes derived from real images. Sato and Cipolla [147] propose semi-local quasi-invariants of curves, which do not need high order derivatives. Nevertheless, their affine quasi-invariants involves second order derivatives. This still is unrealistic for curves extracted from real images, even after a smoothing step. Notice that in the whole book, no second derivative will be involved in the shape recognition, and even not a first derivative (tangents are not used, only bitangents).
Cohen *et al.* [36, 83] propose to approximate curves with B-Splines, leading to a compact representation. This interpolation appears to be robust to noise, and an adequate matching algorithm permits to deal with occlusions. Although this method seems promising, it suffers from the interpolation in itself, which depends on the original sampling of the considered curve.

Most local recognition methods involve curvature extrema of the curves bounding the solid shapes. These points are not affine invariants of curves, but are certainly, from the perceptual viewpoint, the most salient points of shapes. This was already pointed out by Attneave in his 1954 paper [13]:

> *"Information is concentrated along contours (i.e., regions where color changes abruptly), and is further concentrated at those points on a contour at which its direction changes most rapidly (i.e., at angles or peaks of curvature)."*

(See Figure 5.8.) Cohignac *et al.* [38] propose a multiscale curvature representation for shape recognition, by considering curvature extrema of surfaces derived from a shape with the affine morphological scale space. This leads, for each shape, to a set of points of interest in $\mathbb{R}^3$. In such local shape recognition methods, shapes are represented by a finite code, composed by the coordinates of curvature extrema points. Then, recognition can be made local or semi-local by comparing the codes through the partial Hausdorff distance [84]. Two variations based on this general method, leading respectively to a similarity invariant and to a translation-rotation invariant recognition methods, can be found in [8, 65]. Similar approaches can be considered by using the boundary points which are tangent to bitangent lines, instead of the curvature extrema [136].

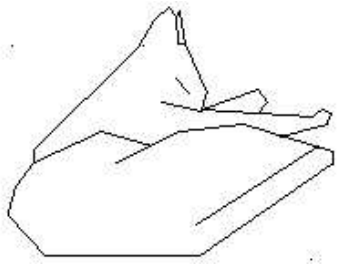


Figure 5.8: (From [13]) Curvature extrema concentrate a large amount of shape information. Quoting Attneave: "Common objects may be represented with great economy, and fairly striking fidelity, by copying the points at which their contours change direction maximally, and then connecting these points appropriately with a straightedge"

Up to here, mainly local invariant features were discussed. Since invariants which are too local, such as differential invariants, suffer from noise, and the global ones (*e.g.* moment invariants) instead suffer from occlusions, a suitable trade-off solution can be the use of semi-local features.
Lamdan *et al.* [99], followed by Rothwell [142, 143], have proposed semi-local descriptors of shapes, invariant up to similarity or affine transforms. (Rothwell *et al.* also propose projective invariant representations.) These features are based on the description of pieces of non-convex curves lying between two bitangent points (*i.e.* points at which the same straight line is tangent to the curve). Such features are affine invariant and the use of bitangent lines ensures robustness to noise. Lisani *et al.* [106, 107] improved this bitangent method by associating, with each bitangent to each level line, a local coordinate system and defining a local affine or similarity normalized piece of curve. They have also added to the representation, similar local invariant descriptions based also on tangent lines to the curve at inflexion points, leading to a more complete representation of level lines.

Some recent methods of image analysis rely on invariant points of interest. These points are singularities of the image related to zero-crossings as in Lowe [110], or to Harris points [77]. By using locally computed affine invariant moments, these points can also be made affine invariant [118]. The purpose is merely to extract an invariant neighborhood of the image, independently of the shape they may contain. However, since interest points are usually located near relevant parts of shapes (see Fig. 5.8), some accurate semi statistical descriptors can be defined. For instance, the descriptors of [110] are local distributions of the gradient direction in some invariant neighborhoods of the points of interest, and are used in [152] for retrieval of image parts in video sequences.

Part III

The recognition of shape elements

Chapter 6

A contrario decision

hile shape comparison, shape matching and shape extraction have been the subject of many researches, the decision step has been rarely studied. This chapter presents a framework to answer by *yes* or *no* the question "does that shape element look like this one?", and to measure the confidence level in this answer. Since the general recognition problem is of interest, any *a priori* information cannot be used. Thus, an *a contrario* model is natural: Two shape elements match if the probability that two shape elements, generated by a suitable a contrario model, are as similar as the ones that are actually observed, is very small. In order to reach high levels of confidence (i.e. small number of false alarms), independent features are extracted from shape elements, such that the model satisfies the Helmholtz principle: any detection in noise should be considered as not relevant. The theoretical value of the number of false alarms gives a good estimate of the number of matches in a white noise image. By imposing that this value is less than 1, there is nearly no false matches.

6.1 *A contrario* models

6.1.1 Shape element model *versus* background model

In what follows, it is always assumed that shape elements have been normalized in the sense of the previous chapter. Let us consider a given query shape element $\mathcal{S}$ and a database $\mathcal{B}$ of N shape elements. Let us also assume that a distance or similarity measure d between shape elements is defined. For $\mathcal{S}' \in \mathcal{B}$, assume that it is observed that $d(\mathcal{S}, \mathcal{S}')$ is "small". (One of the main purposes of the following discussion is to define what "small" does mean.) Such a coincidence may be explained by one of the two following hypotheses:

- $\mathcal{H}_0$: $\mathcal{S}'$ is near $\mathcal{S}$ only *by chance*, for instance because N is very large and the database contains many shape elements.
- $\mathcal{H}_1$: $\mathcal{S}'$ is near $\mathcal{S}$ because of some common causality.

A model for Hypothesis $\mathcal{H}_1$ would be equivalent to a model of shape variability. Accurately defining such a model would require huge sets of observations, which are often not available, or limited to very specific types of shapes. On the other hand, $\mathcal{H}_0$ models random shapes, that are not supposed to match with one another, as for instance shape elements extracted from a white noise image. It will be easier to propose sound models for this hypothesis. A model for $\mathcal{H}_0$ will be called *background model*. This background model is kept formal for the time being, but its construction will be proposed in Sect. 6.2.3.
A classical test for deciding between $\mathcal{H}_0$ and $\mathcal{H}_1$ is to compare the distance $d(\mathcal{S}, \mathcal{S}')$ with some predetermined value δ and to decide that $\mathcal{H}_1$ holds whenever $d(\mathcal{S}, \mathcal{S}') < \delta$. Otherwise, $\mathcal{H}_1$ is rejected and the alternative hypothesis $\mathcal{H}_0$ is accepted. The quality of a statistical test is measured by the probability of taking wrong decisions. Two kinds of errors are possible: reject $\mathcal{H}_1$ for an observation $\mathcal{S}'$ for which $\mathcal{H}_1$ is actually true (non-detection or type I error), and accept $\mathcal{H}_1$ for $\mathcal{S}'$ although $\mathcal{H}_1$ is false (false alarm, or type II error). A probability measure can be associated to each type of error.

- the *probability of non-detection* or *probability of a miss* (associated with type I error)

$$PM(\mathcal{S},\delta) \equiv \Pr(d(\mathcal{S},\Sigma) \geqslant \delta | \mathcal{H}_1);$$

- the *probability of false alarms* (associated with type II error)

$$PFA(\mathcal{S},\delta) \equiv \Pr(d(\mathcal{S},\Sigma) < \delta | \mathcal{H}_0), \tag{6.1}$$

provided $\Pr(\cdot)$ is a probability measure defined on the set of shape elements. Here and thereafter, we adopt the convention to use uppercased Greek letters for random shape elements, while Roman letters will be used for observed values.

It is worth noticing that the background model is given by $\Pr(\cdot|\mathcal{H}_0)$. It is clear that the lower PM and PFA, the better the test, but it is also clear that PM and PFA cannot be independently optimized. The usual problem is to find a trade-off between these two probabilities.
Widely used techniques as *Bayesian test* or the *Neyman-Pearson* test often amounts to threshold the likelihood ratio of the observation under $\mathcal{H}_0$ and $\mathcal{H}_1$ [141]. However, the practical limits of this theoretical framework are obvious. They indeed require the knowledge of the likelihoods of the hypothesis $\mathcal{H}_1$ and the counter-hypothesis $\mathcal{H}_0$. This is in general unrealistic if the aim is to recognize an unspecified query shape element. A generative model would indeed be needed for the query shape element $\mathcal{S}$ if the likelihood of each different shape element Σ under hypothesis $\mathcal{H}_1$ was to be computed. In the Bayesian approach, it is also required and generally not possible to accurately compute the probability of non-detection $\Pr(d(\mathcal{S},\Sigma) \geqslant \delta|\mathcal{H}_1)$. This probability indeed relies on an observation model. (Noise, blur, etc.) Such a model is possible in particular applications, where there are hypotheses on the shapes to be sought for. No such assumption is made in the present context. If two images have shapes in common, these shapes appear in very few instances, and classical methods do not allow to construct models from these samples. On the other hand, it may be easier to model the probability of false alarm $PFA(\mathcal{S},\delta)$, since $\mathcal{H}_0$ models a completely random (i.e. with no information) situation. Our claim is that it is possible to take a decision, only based on the background probability model for $\mathcal{H}_0$. A sure detection simply requires that this probability be very small. In Section 6.2.2, it is explained how to attain such probabilities.

6.1.2 A detection terminology

Since a probability has little meaning *per se*, let us now introduce the *number of false alarms*. Let N denote the number of shape elements in the database.

DEFINITION 6.1 *The* Number of False Alarms *of the shape element $\mathcal{S}$ at a distance δ is*

$$NFA(\mathcal{S},\delta) \equiv N \cdot PFA(\mathcal{S},\delta), \tag{6.2}$$

where $PFA(\mathcal{S},\delta)$ is defined in (6.1).

The number of false alarms is the expected number of the shape elements in the database whose distance to $\mathcal{S}$ is below δ, when it is assumed that $\mathcal{B}$ obeys the background model.

By abuse of language, the quantity $NFA(\mathcal{S}, d(\mathcal{S},\mathcal{S}'))$ is called the *number of false alarms between a query shape $\mathcal{S}$ and a database shape $\mathcal{S}'$*. Therefore, this NFA corresponds to the expected number of shapes in $\mathcal{B}$ that lie closer to the query shape $\mathcal{S}$ than $\mathcal{S}'$ does, when $\mathcal{B}$ has been generated by the background model.

DEFINITION 6.2 *A shape element $\mathcal{S}'$ is an* ε-meaningful match *of the query shape element $\mathcal{S}$ if*

$$NFA(\mathcal{S}, d(\mathcal{S},\mathcal{S}')) \leqslant \varepsilon. \tag{6.3}$$

Considering ε-meaningful matches as pertinent detections is an *a contrario* decision since the hypothesis "a database shape element $\mathcal{S}'$ shares some causality with the query $\mathcal{S}$ " is accepted as soon as it is not likely that $\mathcal{S}'$ has been generated by the background model. The above definition is justified by the following result.

PROPOSITION 6.1 *Under the assumption that the database shape elements are identically distributed following the background model, the expectation of the number of ε-meaningful matches with $\mathcal{S}$ is less than ε.*

Proof: Let Σ_j ($1 \leqslant j \leqslant N$) denote the shape elements in the database, and χ_j the indicator function of the event e_j: "Σ_j is an ε-meaningful match of the query $\mathcal{S}$" (*i.e.* its value is 1 if Σ_j actually is an ε-meaningful match of $\mathcal{S}$, and 0 otherwise). Let $R = \sum_{j=1}^{N} \chi_j$ be the random variable representing the number of shape elements ε-meaningfully matching $\mathcal{S}$.
The key point is that the linearity of the expectation allows to compute $\mathbb{E}_{\mathcal{H}_0}(R)$, the expectation of R in the background model, while it is difficult or impossible to estimate the probability law of R (even under $\mathcal{H}_0$), because of the unknown dependencies between the events e_j. Linearity yields $\mathbb{E}_{\mathcal{H}_0}(R) = \sum_{j=1}^{N} \mathbb{E}_{\mathcal{H}_0}(\chi_j)$. By definition of χ_j,

$$\mathbb{E}_{\mathcal{H}_0}(\chi_j) = \Pr(\Sigma_j \text{ is an } \varepsilon\text{-meaningful match of } \mathcal{S} | \mathcal{H}_0).$$

By definition, Σ_j is an ε-meaningful match of $\mathcal{S}$ if

$$\Pr(\Sigma' \in \mathcal{B},\, d(\mathcal{S}, \Sigma') < d(\mathcal{S}, \Sigma_j) | \mathcal{H}_0) \leqslant \frac{\varepsilon}{N}. \tag{6.4}$$

Remark that the probability on the left-hand side is itself a random variable. The probability of this event is less than $\frac{\varepsilon}{N}$. Indeed, let us denote by X_j the random variable $d(\mathcal{S}, \Sigma_j)$, and F the repartition function (under $\mathcal{H}_0$) of $d(\mathcal{S}, \Sigma)$. Hence, F is also the repartition function of X_j and the event on the left-hand side of (6.4) also reads $F(X_j) < \frac{\varepsilon}{N}$. Lemma 3.2 then implies that

$$\Pr\left(F(X_j) < \frac{\varepsilon}{N} | \mathcal{H}_0\right) \leqslant \frac{\varepsilon}{N}.$$

This yields

$$\mathbb{E}_{\mathcal{H}_0}(R) \leqslant \sum_{j=1}^{N} \frac{\varepsilon}{N} = \varepsilon. \qquad \square$$

This methodology does not enable to estimate *a priori* the number of ε-meaningful matches in a database of shape elements extracted from a natural image (i.e. whose shape elements are not likely generated by the background model.) This number is an output of the method. The idea behind the definition is that if all shape elements in the database were generated by the background model, then Hypothesis $\mathcal{H}_1$ should never be accepted. In this case, all ε-meaningful detections should be considered as false alarms. In average, there are less than ε detections.

The lower ε, the "surer" the ε-meaningful detections. Of course, the same claim is true when considering distances: the lower the distance threshold δ, the surer the corresponding matches, but considering the NFA quantifies this confidence level. Actually, by monotonicity, the equation

$$\delta^*\left(\frac{\varepsilon}{N}\right) \equiv \sup\{\delta > 0, PFA(\mathcal{S}, \delta) \leqslant \varepsilon/N\}$$

suitably defines a positive real number. The proposition that follows is then straightforward.

PROPOSITION 6.2 *A shape element $\mathcal{S}'$ is an ε-meaningful match of the query $\mathcal{S}$ if and only if $d(\mathcal{S}, \mathcal{S}') < \delta^*\left(\frac{\varepsilon}{N}\right)$.*

Thus, selecting ε-meaningful matches is equivalent to selecting $\mathcal{S}'$ such that $d(\mathcal{S}, \mathcal{S}') < \delta^*\left(\frac{\varepsilon}{N}\right)$. In practice, the method consists in fixing ε, and the value $\delta^*\left(\frac{\varepsilon}{N}\right)$ remains implicit. Moreover, computing the NFA does not need any shape model for $\mathcal{S}$.

The definition of the number of false alarms can be extended to the case of the comparison of all the shape elements of two databases.

DEFINITION 6.3 *Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two databases containing respectively N_1 and N_2 shape elements. The* Number of False Alarms *of a shape element $\mathcal{S}$ (belonging to $\mathcal{B}_1$) at a distance δ is*

$$NFA(\mathcal{S}, \delta) = N_1 \cdot N_2 \cdot PFA(\mathcal{S}, \delta). \tag{6.5}$$

This situation corresponds to experiments in Chap. 7 where the shape contents of pairs of images are compared. Prop. 6.1 then remains true, that is to say if $\mathcal{B}_2$ is generated by the background model, the expected number of ε-meaningful matches between $\mathcal{B}_1$ and $\mathcal{B}_2$ is less than ε.

6.1.3 Recognition threshold is relative to the context

Notice that the empirical probabilities take into account the "rareness" or "commonness" of a possible match. Indeed the threshold δ^* will be large in the first case and small in the other one. If a query shape element $\mathcal{S}_1$ is less common than another one $\mathcal{S}_2$, then the database contains more shape elements close to $\mathcal{S}_2$ than shape elements close to $\mathcal{S}_1$. As a consequence,

$$PFA(\mathcal{S}_1, \delta) \leqslant PFA(\mathcal{S}_2, \delta).$$

Another important issue is the role of the empirical learning of the probability distribution PFA. When searching for $\mathcal{S}$ in $\mathcal{B}$, the law of $d(\mathcal{S}, \cdot)$ is empirically learned (see Sect. 6.2.3). Usually, $\mathcal{B}$ is used, but another database $\mathcal{B}_2$ could be used instead for learning the probabilties. This then yields two different distance thresholds. The influence of this phenomenon will be empirically studied in the next chapter, in Sect. 7.2.

6.2 Why an *a contrario* decision?

6.2.1 Controlling the number of false alarms with no *a* priori

The advantages of the *a contrario* decision framework based on the NFA compared to directly setting a distance threshold between shape elements are clear. Simply setting $\varepsilon = 1$ allows at most one false alarm among meaningful matches (1-meaningful matches are also simply refered to as "meaningful matches"). In practice, setting $\varepsilon = 10^{-1}$ eliminates all false detections. The detection threshold ε is set uniformly whatever the query shape element and the database may be: the resulting distance threshold adapts automatically according to them as explained in the preceding section. On the other hand, the lower ε, the "surer" the ε-meaningful detections are. Of course, the same claim is true when considering distances: the lower the distance threshold δ, the surer the corresponding matches, but considering the NFA quantifies this confidence level.

6.2.2 How to attain very small numbers of false alarms?

Consider the following heuristic argument. Assume that the distribution of $d(\mathcal{S}, \Sigma)$ is learned by empirical frequencies on a set of N shape elements. Then, the lowest non null observable probability is $1/N$. If $\mathcal{S}$ is now sought for in another database also containing N shape elements, then the lowest attainable number of false alarms is $N \cdot \frac{1}{N} = 1$. This means that even if two shape elements $\mathcal{S}$ and $\mathcal{S}'$ are almost identical, such an empirical estimate of the NFA cannot ensure that this match is not casual. Indeed, an NFA equal to 1 means that, on the average, one of the shape elements in the database can match $\mathcal{S}$ by chance. In his pioneering work, D. Lowe [111] presents this same viewpoint for visual recognition:

> *Due to limits in the accuracy of image measurements (and possibly also the lack of precise relations in the natural world) the simple relations that have been described often fail to generate the very low probabilities of accidental occurrence that would make them strong sources of evidence for recognition. However, these useful unambiguous results can often arise as a result of combining tentatively-formed relations to create new compound relations that have much lower probabilities of accidental occurrence.*

Let us make this remark more specific. Assume that each shape element $\mathcal{S}$ can be represented by a set of K features $x_1(\mathcal{S}), \ldots, x_K(\mathcal{S})$, each of them belonging to a metric space (E_i, d_i) ($i \in \{1, \ldots, K\}$). Let us denote by $P_i(\mathcal{S}, \delta)$ the marginal probability

$$P_i(\mathcal{S}, \delta) = \Pr(d_i(x_i(\mathcal{S}), x_i(\Sigma)) \leqslant \delta | \mathcal{H}_0). \tag{6.6}$$

Let us also make the assumption, that, in the background model, the features $x_i(\Sigma)$ are mutually independent variables. Actually, this independence assumption can be taken as a definition of the background model, as follows.

DEFINITION 6.4 *A* background model *is a probability system such that the following assumption holds:*
(A) *The random variables* $\Sigma \mapsto d_i(x_i(\mathcal{S}), x_i(\Sigma))$ *(*$i \in \{1, \ldots, K\}$*) are mutually independent.*

From the partial distances d_i, a complete, global distance should be defined, in order to apply the results of Sect. 6.1.1. A possible choice could be the product distance d defined by

$$d(\mathcal{S}, \mathcal{S}') = \max_{i \in \{1, \ldots, K\}} d_i(x_i(\mathcal{S}), x_i(\mathcal{S}')). \tag{6.7}$$

Nevertheless, there is no reason why the d_i should have the same order of magnitude. Instead, let us define

$$\delta_i(\mathcal{S}, \mathcal{S}') = P_i(\mathcal{S}, d_i(\mathcal{S}, \mathcal{S}')) \tag{6.8}$$

i.e.

$$\delta_i(\mathcal{S}, \mathcal{S}') = \Pr(d_i(x_i(\mathcal{S}), x_i(\Sigma)) \leqslant d_i(x_i(\mathcal{S}), x_i(\mathcal{S}') | \mathcal{H}_0). \tag{6.9}$$

Let us also define the product distance

$$d(\mathcal{S}, \mathcal{S}') = \left(\max_{i \in \{1, \ldots, K\}} \delta_i(\mathcal{S}, \mathcal{S}') \right)^K. \tag{6.10}$$

Remark that, despite the denomination, this function is not necessarily a distance. However, $d(\mathcal{S}, \mathcal{S}')$ is small when observing random values $d_i(x_i(\mathcal{S}), x_i(\Sigma))$ smaller than $d_i(x_i(\mathcal{S}), x_i(\mathcal{S}'))$ occurs with a low probability. Hence d is a measure of dissimilarity which is relative to $\mathcal{S}$.

The purpose of this operation is the following: if Σ is now a random shape element, $\delta_i(\mathcal{S}, \Sigma)$ is also a random variable. If the distribution of the distances d_i are given by densities, then δ_i is uniform in $(0, 1)$ (by Lem. 3.2), whatever the law of Σ. Of course, the δ_i are independent if the d_i are independent, which is assumed in the background model.

The NFA between $\mathcal{S}$ and $\mathcal{S}'$ is still defined by

$$NFA(\mathcal{S}, \mathcal{S}') = N \cdot d(\mathcal{S}, \mathcal{S}').$$

The next result immediately generalizes Prop. 6.1.

COROLLARY 6.1 *The expected number of* ε*-meaningful matches in a database of* N *shape elements generated by the background model is less than* ε*.*

Proof: The sketch of the proof follows the one of Prop. 6.1. By linearity of the expectation, it suffices to prove $\Pr(NFA(\mathcal{S}, \Sigma) < \varepsilon | \mathcal{H}_0) < \frac{\varepsilon}{N}$. By definition, $NFA(\mathcal{S}, \Sigma) < \varepsilon$ if and only if, for all $i \in \{1, \ldots, K\}$,

$$\delta_i(\mathcal{S}, \Sigma) = P_i(\mathcal{S}, d_i(\mathcal{S}, \Sigma)) < \left(\frac{\varepsilon}{N} \right)^{1/K}.$$

By the independence assumption,

$$\Pr(NFA(\mathcal{S}, \Sigma) < \varepsilon | \mathcal{H}_0) = \prod_{i=1}^{K} \Pr\left(P_i(\mathcal{S}, d_i(\mathcal{S}, \Sigma)) < \left(\frac{\varepsilon}{N} \right)^{1/K} | \mathcal{H}_0 \right).$$

But since P_i is exactly the repartition function of $d_i(\mathcal{S}, \Sigma)$, Lem. 3.2, (p. 17) applies and each probability on the right hand product is less than $\left(\frac{\varepsilon}{N}\right)^{1/K}$. Hence,

$$\Pr(NFA(\mathcal{S}, \Sigma) < \varepsilon | \mathcal{H}_0) \leqslant \frac{\varepsilon}{N}.$$

This allows to conclude. □

Consider again the numerical heuristic argument above, where the $P_i(\mathcal{S}, \delta)$ are empirically learned on a size N database. The smallest attainable number of false alarms is $N \cdot \frac{1}{N^K} = N^{1-K}$ which can now be much less than 1. Note that the independence assumption is not unrealistic under hypothesis $\mathcal{H}_0$ while it may certainly not be true under $\mathcal{H}_1$. This also leads to the following consideration. The number of features K should be large enough, so as to attain very small NFAs. But, it cannot be arbitrary large either. Indeed, digital images contain a finite amount of information. Therefore, even in a white noise image, shape elements cannot be described by a infinite set of independent features. In particular, the features x_i must be measured at some locations that are afar at least by the Nyquist distance, before normalization. On the other hand, the features $x_i(\mathcal{S})$ should characterize a shape $\mathcal{S}$, so that $d(\mathcal{S}, \mathcal{S}')$ is small if and only if $\mathcal{S}$ and $\mathcal{S}'$ are similar. Finding a suitable trade-off between independence and completeness of the features is necessary.

6.2.3 Deriving statistically independent features from shape elements

The decision framework described so far is actually completely general, in the sense that it can be applied to find correspondences between any kind of structures for which K statistically independent features can be extracted. Let us now concentrate on the problem of extracting independent features from normalized shape elements. In order to make the shape recognition task reliable, shape features have to meet the three following requirements:

1) Completeness: two shape elements are alike if and only if their features are alike.

2) Statistical mutual independence (more precisely, distances between features are independent).

3) Their number is as large as possible.

The first requirement means that the features describe shapes well, the second one is imposed in order to design the background model, and the third requirement is needed in order to reach low numbers of false alarms. The existence of a background model as defined in Def. 6.4 is not obvious. In particular, proving independence is not an easy task. The remainder of this section describes a possible construction of shape element features, both in the semi-local and global cases.

Semi-local encoding

Let us first consider the semi-local encoding algorithm described in Chapter 5. Recall that a normalized shape element is in fact a piece of Jordan curve, normalized in a local frame built on a bitangent or on a flat part. The construction to be described now was empirically found to yield a good trade-off achieving simultaneously the three feature requirements (see Fig. 6.1 for an illustration). Each normalized representation C is split into five pieces of equal length. Each one of these pieces is normalized by mapping the chord between its first and last points onto the horizontal axis, the first point being at the origin: the resulting normalized chunks are five features C_1, C_2, ...,C_5. These features ought to be independent; nevertheless, C_1, ...,C_5 being given, it is impossible to reconstruct the shape element they come from. For the sake of completeness a sixth, global, feature C_6 is therefore made of the endpoints of the five previous pieces, in the normalized frame. For each piece of level line, the shape features introduced in section 6.2.2 are made of these six "generic" shape codes C_1, ...,C_6. Using the notations introduced in the previous sections, $x_i(\mathcal{S}) = C_i$, $i \in \{1, \ldots, 6\}$; the distances d_i between them are L^∞-distances between corresponding pieces, parameterized by length. If $C_i(s)$ is such a parameterization, let us simply set $d_i(C_i(s), \tilde{C}_i(s)) = \sup_s \|C_i(s) - \tilde{C}_i(s)\|$.

The independence hypothesis amounts to say that, for shapes of the reference database, a part of a shape element does not influence the other ones, and that scales do not interfere, which is not too strong an assumption. Whereas it cannot be proved that this description provides a background model in the sense of Def. 6.4, its consistency with the theory of Sect. 6.1 will be empirically checked.

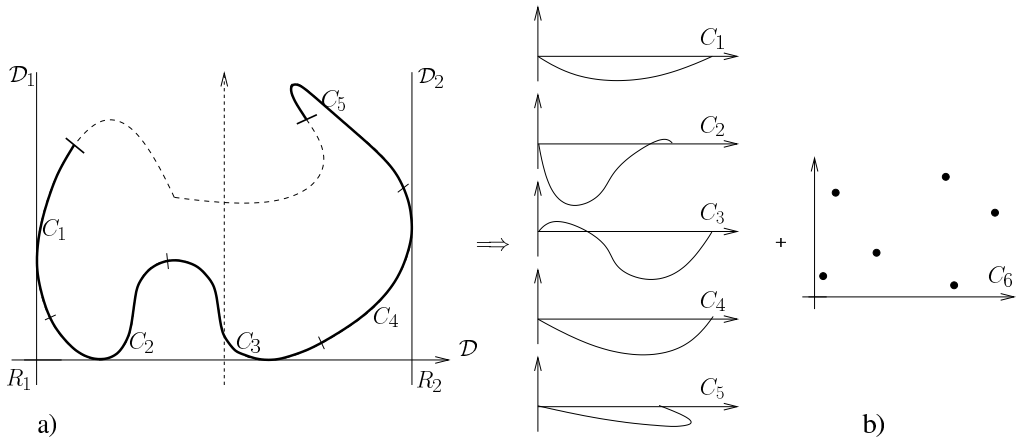


Figure 6.1: Semi-local encoding procedure. Example of a similarity-invariant encoding. Sketch (a): original shape element in a normalized frame based on a bitangent line. Both ends of the piece of the shape element, of length proportional to $\|R_1R_2\|$, are marked with bold lines: this representation is split into 5 pieces C_1, C_2, C_3, C_4, and C_5. Sketch (b): each of them is normalized, and a sixth feature C_6 made of the endpoints of these pieces is also built

Let us give some realistic orders of magnitude. In typical 512×512 images, the experimental number of extracted shape elements is about 10^4. Thus, the smallest number of false alarms when matching shape elements between two images is

$$10^4 \cdot 10^4 \cdot \frac{1}{(10^4)^6} = 10^{-16}.$$

In practice, for similar shapes, numbers of false alarms as small as 10^{-10} are observed.

Global encoding

A global curve normalization was also proposed in Chapter 5 . The *a contrario* decision strategy is still valid, considering these normalized curves as shape elements, and building the features in a similar way as for the semi-local encoding. Precisely speaking, each normalized piece of curve is split into five pieces. The starting point was defined in Chapter 5 as the nearest point to the barycenter intersecting the vertical line to the bottom, with a positive ordinate. In the same way as for semi-local encoding, each one of these pieces is normalized by mapping the chord between its first and last points onto the horizontal axis, the first point being at the origin: the resulting chunks are five features C_1, C_2, ...,C_5. For the sake of completeness, a sixth global feature C_6 is made of the endpoints of the five previous pieces. The features are made of C_1, ...,C_6. The distances d_i between them are again L^∞-distances between corresponding chuncks, parameterized by length.

6.3 Testing the background model

The computation of the probability $PFA(\mathcal{S}, \delta)$ that a shape element could be just by chance at a distance lower than δ to $\mathcal{S}$ is correct under the independence assumption on the pieces of shape elements. Of course, the degree of trust that can be given to the associated Number of False Alarms $NFA(\mathcal{S}, \delta)$ (Definitions 6.1, and 6.3) strongly depends on the validity of this independence assumption. Before applying this method to realistic

data, the independence of the pieces of shape elements must be tested, in order to ensure the correctedness of the methodology. This is the aim of what follows. Although a decision rule is proposed for both global and semi-local shape elements matching, only the results of the tests for semi-local encoding are given. It is shown here that the pieces of shape elements obtained by the proposed normalization (section 6.2.3) are *not* independent (section 6.3.1), as some dependence is introduced by the non-self intersection constraint of level lines and (mainly) by the normalization procedure (section 6.3.2). Nevertheless, experiments point out that detection under Helmholtz principle (*i.e.* a meaningful match is a match that is not likely to occur in a noise image) is fully satisfying (section 6.3.2).

6.3.1 Independence testing

In order to compute the probability $PFA(\mathcal{S}, \delta)$, the mutual independence of the "chunks of shape elements" is needed. More precisely speaking, a shape element made of pieces x_i being given, the binary random variables $y \mapsto d_i(x_i, y) \leqslant \delta$ are supposed mutually independent in a situation where it is *a priori* known that the shapes are different, typically white noise.
It is not possible to estimate the joint probability

$$\Pr\left(\Sigma,\, d_1(x_1, x_1(\Sigma)) \leqslant \delta, \ldots, d_n(x_n, x_n(\Sigma)) \leqslant \delta\right).$$

and compare it to the product $\prod_{i=1}^{n} \Pr(\Sigma,\, d_i(x_i, x_i(\Sigma) \leqslant \delta)$. (Estimating the law of this random vector would indeed require too many samples.) On the other hand, it turns out that the joint probability associated to two pieces of shape elements can be accurately enough estimated. Thus, the mere pairwise independence is tested, instead of the mutual independence of the pieces of shape elements. A classical Chi-square test shows that the pairwise independence does not hold. The results are clear: in all cases, the independence assumption is rejected with a high significance level. Nevertheless, the rejection is strong because the tested databases are very large: Chi-square test is all the more accurate (and so is the rejection confidence) as the number of samples is large. In other terms, a "slight" dependence with a large number of samples may lead to a very significant rejection; this means that the Chi-square test does not yield an absolute measurement of how dependent or how independent variables are.
The next section shows that the independence assumption is true enough to keep the Helmholtz detection principle true, in a sense that will be made clear.

6.3.2 Checking the Helmholtz principle

The purpose of this section is to test the main property of the proposed method, namely the validity of the computation of the expected number of "random" detections (the Number of False Alarms, Prop. 6.1). This computation holds under an independence assumption. If this assumption is true, and if the database contains no copy of a sought for shape element (*i.e.* the query shape element is not "generic" among the shape elements in the database), the expected number of false alarms among all ε-meaningful matches with the query shape element should be lower than ε. Nevertheless, false alarms and real matches cannot be distinguished *a priori*: they all are ε-meaningful detections. The Chi-square test proved that, strictly speaking, the independence assumption is not valid. Now, Helmholtz principle states that no detection in "noise" (whose model has to be precised) should be considered as relevant. All ε-meaningful matches in noise should thus be considered as false alarms: in such a situation there should be on the average about ε many of them. The following experiments test this claim. It turns out that the NFA is a good prediction of the number of detections. The independence assumption is valid enough, so that the claim according to which there is on average at most ε false alarms among ε-meaningful matches still holds.

As a first experiment, let us check the detection thresholds on a very simple model: the shape element database and shape element query are random walks with independent increments. In this case the background model is ensured to be true, in the sense that the considered shape elements perfectly fit the independence assumption. Table 6.1 shows that the Number of False Alarms is very accurately predicted for various database

Table 6.1: Random walks. Average number of detections (over 10 samples) vs ε. The experiments were made with databases of different size (N from 10,000 to 100,000 shape elements)

N \ ε	0.01	0.1	1	10	100	1,000	10,000
100,000	0	0	2.3	15.2	122.2	1,075.5	9,872.2
50,000	0.2	0.3	1.5	11.9	106.1	1,001.1	9,789.5
10,000	0	0	1.2	12.5	108.4	985.0	—

Table 6.2: Pieces of white noise level lines with *no* normalization. Average (over 10 samples) number of detections vs ε. The three rows correspond to the size of the various databases, respectively $N = 101,438$ shape elements, $50,681$ shape elements and $9,853$ shape elements

N \ ε	0.01	0.1	1	10	100	1,000	10,000
101,438	0.1	0.1	1.7	13.8	95.3	942.5	9,789.4
50,681	0	0	1.2	10.3	90.5	955.1	9,859.3
9,853	0	0.1	0.9	9.5	94.3	973.1	—

sizes: the number of detections with a NFA lower than ε is about ε indeed. Of course, modelling shape elements with random walks is not realistic. As proved in what precedes, distances between shape elements are actually *not* independent. This lack of independence may come from two different sources. On the one hand, shape elements correspond to pieces of level lines, and consequently are constrained not to self-intersect. On the other hand, shape elements are normalized, and show therefore structural similarities (for example, shape elements coming from bitangent points show mostly common structures).

In order to quantify the "amount of dependence" due to these two aspects, let us consider the two following experiments. Table 6.2 shows the number of detections *versus* the number of false alarms for databases made of pieces of level lines (*not* normalized, the shape elements are just made out of 45 consecutive points on pieces of level lines). Consequently, the obtained shape elements are constrained not to self-intersect. In this experiment, the independence can only be spoiled by this property, not by the normalization. Although the Chi-square test shows that the shape elements are not independent, once again the number of detections is accurately predicted: the number of matches with a NFA less than ε is indeed about ε. Let us consider databases made of normalized shape elements extracted from pieces of level lines in white noise images. Table 6.3 shows that the number of detections is still of the same magnitude as the number of false alarms ε, but is not as precisely predicted as in the latest experiments. Roughly speaking, this means that the dependence "mostly comes from the normalization procedure", and not from the non-self-intersection constraint. Nevertheless, the order of magnitude is still correct, and does not depend on the size of the database. These properties are sufficient for setting the Number of False Alarms threshold under the Helmholtz principle. Following this method, a match is supposed to be highly relevant if it cannot happen in white noise images. According to table 6.3, matches with a NFA lower than 0.1 are ensured to be unlikely in white noise images. Requiring a strong confidence in the detected matches thus leads to consider 0.1-meaningful matches in realistic experiments (see Chap. 7, Sect. 7.1). As a last experiment, Table 6.4 shows the number of detections *versus* number of false alarms for a database made of normalized long (length greater than 135 pixels) pieces of level lines from white noise images. The results are not better than in the preceding experiment, so that it cannot be asserted that the independence violation is due to short pieces of level lines.

Table 6.3: Normalized pieces of white noise level lines. Average (over 10 samples) number of detections vs ε on databases with respective size $N = 104,722$, $47,033$ and $10,784$ shape elements

N \ ε	0.01	0.1	1	10	100	1,000	10,000	100,000
$104,722$	0.3	1.5	6.5	31.5	173.9	$1,264.4$	$9,803.1$	$99,899.5$
$47,033$	0.1	0.3	3.7	20.2	125.4	976.3	$9,854.2$	—
$10,784$	0	0.2	2.6	14.8	107.6	973.3	—	—

Table 6.4: Normalized long (more than 135 pixels) pieces of white noise level lines. Average (over 10 samples) number of detections vs ε, for different size of databases N (indicated in the left column)

N \ ε	0.01	0.1	1	10	100	$1,000$	$10,000$	$100,000$
$101,743$	0	0.4	2.8	18.5	124.3	$1,123.2$	$9,693.8$	$99,921.0$
$51,785$	0	0.3	2.9	16.0	118.6	983.4	$9,800.4$	—
$11,837$	0	0.2	1.4	12.3	105.9	975.2	$9,974.7$	—

6.4 Bibliographical notes

6.4.1 Shape distances

The shape matching problem is strongly related to the definition of adequate distances. The most commonly used distances are L^p distance, Mahalanobis distance [56, 154], Hausdorff distance [84], or Fréchet distance [4]. M. Miller, L. Younes and A. Trouvé [119, 120] (see also the more recent [18]) study the orbit of shapes via the action of diffeomorphic transformations, allowing by this way non-rigid transformations. Each transformation has a cost, and the distance between two shapes is the cost of the transformation of least energy between them. Similar distances as cost of elastic deformations have been elaborated by [17]. Most of these distances are global and sensitive to local occlusions; however, they can be suitably modified to fit the locality requirement, leading for instance to partial Hausdorff distance [84, 158, 137]. We refer the reader to general surveys by Alt et al. [3], Veltkamp et al. [158, 160], Loncarnic [109] and Dryden [53]. A review of more applied methods, involved in "Content-based image retrieval systems", can be found in [159, 161].
Some global features allow to match shapes based on other criteria than invariance with respect to a projective subgroup. For instance, a lot of work has been done on methods for matching shapes by minimizing the deformation energy involved in aligning one shape with another. One such method is modal matching [149], which takes a certain physical plausibility of the deformations into account, and thus accepts a larger class of invariance than geometric groups. Methods minimizing non-rigid energy deformations can also be based on local features, but they do not allow partial matching, since all features are involved in the deformation energy. As an example, Belongie et al. [20] propose to estimate the transformation leading from one shape to another, when each shape is described by some points with a "shape context" (information about the points vicinity). Lisani et al. [106, 107] first defined shape elements as pieces of level lines. The present normalization is basically the same, but distance thresholds are automatically determined, while they were empirically chosen by Lisani.

6.4.2 A contrario methods

A contrario detection frameworks are classical in the signal processing field, where a precise model of noise is often available. See for example an application to the detection of gravitational burst in Arnaud et al. [9], and

another to the detection of small targets in cluttered environment in Chapple et al. [35]. In both cases, in the absence of signal, the data distribution is assumed to be a zero mean gaussian with known variance.
An example of target detection in non-Gaussian images can be found in Watson and Watson [163]. The authors model the background of the considered images with a fractal model based on a wavelet analysis. Targets are detected as rare events with regard to this model.
The a contrario detection framework has recently been applied by Desolneux et al. to the detection of alignments [47] or contrasted edges [48], by Almansa et al. to the detection of vanishing points [2], by Stival and Moisan to stereo images [123], by Gousseau to the comparison of image "composition" [71] and by Cao to the detection of good continuations [29].
Another possibility that was investigated is to use the principal component analysis (PCA) [132]. Although PCA does not provide independent features but uncorrelated ones, the approximation does not seem to be critical. However, the completeness requirement (for the same number of features) is not satisfied with PCA. Moreover, there is no good reason why shape elements form a vector space. The same remark holds for independent component analysis (ICA) [86], which assumes that the "signals" (here, shape elements) are linear mixtures of independent features.

Chapter 7

Experiments on meaningful matches detection

his chapter presents experiments illustrating the method for identifying shape elements that was described in the previous chapter. Section 7.1 deals with the semi-local invariant recognition method. Both similarity and affine methods are considered, and a comparative study based on some examples is presented. When images differ by a similarity, affine matching usually returns less matches, because affine encoding is more demanding. Nevertheless, it obviously proves more robust as soon as there is a slight perspective effect. In this case, NFA are much lower than in the similarity invariant case. Global matching is also accurate while unsufficient on its own: many convex shapes can be proved equivalent up to an affine transformation. Finally, the relativity of recognition with respect to the context is illustrated.

Now comes the time to check the applicability of the shape comparison scheme described in the previous chapters. All the experiments presented thereafter follow the same procedure: detection of meaningful boundaries (Chap. 3), affine invariant smoothing (Chap. 4.3), similarity or affine normalization and encoding (Chapters 4 and 5), then matching (Chap. 6).

7.1 Local meaningful matches

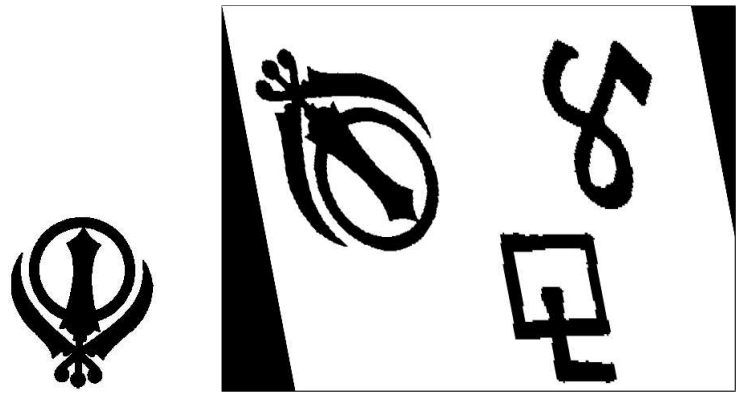
This section presents several experiments that illustrate all the stages of the semi-local invariant recognition method, in particular the semi-local normalization procedures (Chap. 5) and the decision method (Chap. 6). Both similarity and affine versions are considered and compared.

7.1.1 Toy example

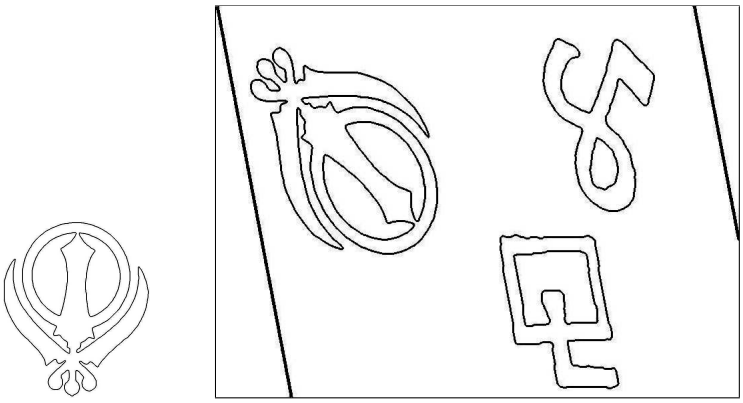
This first experiment compares the performance of the affine invariant and the similarity invariant recognition methods, on simple, synthetic images. A toy example was chosen here in order to illustrate all the stages in the considered recognition methods. Figure 7.1 shows the two synthetic images involved in the experiment. Shape elements from the image on the left (the "query" image) are searched in the right image (the "scene" image).

In the scene image, an affine distorted version of the symbol in the query image is included. The affine and the similarity semi-local invariant encoding algorithms, described in Chap. 5, were applied to the smoothed extracted boundaries, before meaningful matches were detected in both cases.

Let us start by giving some details on the semi-local affine invariant recognition method, and by describing its results. In the query image, 44 shape elements were extracted from its meaningful boundaries. These shape elements are represented by affine normalized codes of 45 points, as explained in Chap. 5. The same encoding procedure, applied to the scene image, led to 105 normalized shape elements. Meaningful matches between these two sets of normalized shape elements were detected. Following the rationale for the meaningfulness computation presented in Chap. 6, a perfect match between shape elements would have reached a NFA of



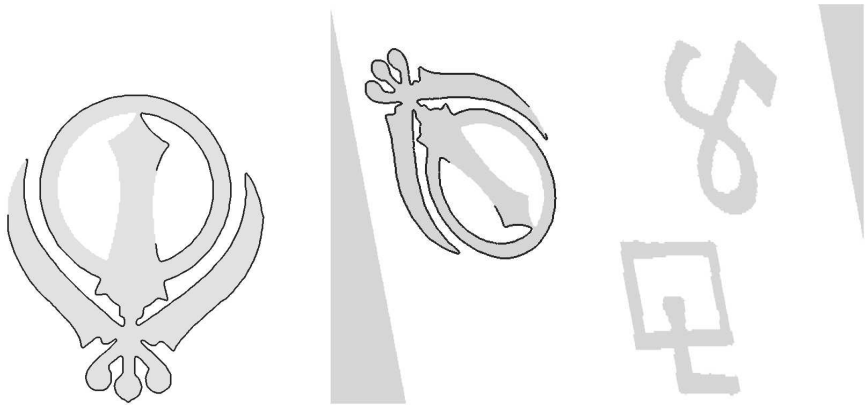
(a)



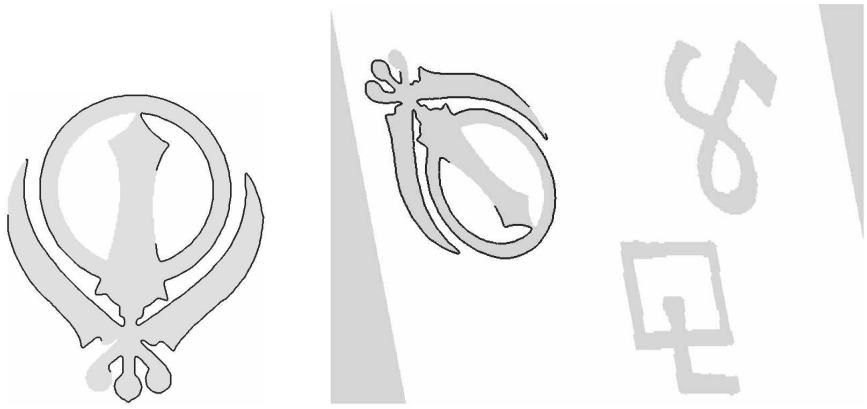
(b)

Figure 7.1: Toy example. (a) Original images; the image on the right contains an affine distorted version of the symbol in the left image. (b) Corresponding maximal meaningful boundaries

$44 \times 105/105^6 = 3.45\,10^{-9}$ (when the empirical distributions of distances to query shape elements are learned using only the considered scene image, as done here). But perfect matching is impossible even in this condition, where synthetic images are considered. This is due to the fact that the interpolation involved in the affine transformation of the image leads to boundaries that are not exactly the transformed boundaries of the original image. Another reason is that, as pointed out in Chap. 5, flat parts are not affine invariant (they are not even similarity invariant), and their position may vary, particularly when dealing with curves showing relatively high curvature. This is exactly what can be observed in this experiment. All 42 detected meaningful matches between shape elements for the affine invariant framework ($NFA < 1$) are shown (superimposed) in Figure 7.2(a). No false match was detected. The best match attains $NFA = 5.4\ 10^{-7}$, and the worst one has an NFA of $9.6\ 10^{-1}$. These two matches are displayed in Figure 7.3(a); the leftmost and middle images correspond respectively to the query and the scene shape elements, and the rightmost image shows their normalized shape elements in the normalized frame, superimposed. The shape elements matching at $NFA = 9.6\ 10^{-1}$ do not correspond exactly to the same piece of curve, but they are still detected since they are relatively close. This kind of instability is not really a problem, since in general the encoding is redundant enough to capture better matches involving the same portions of the curve. This is illustrated in Figure 7.2(b), where almost all the same pieces of boundary shown in Figure 7.2(a) are still present with a meaningfulness $\varepsilon < 10^{-2}$.



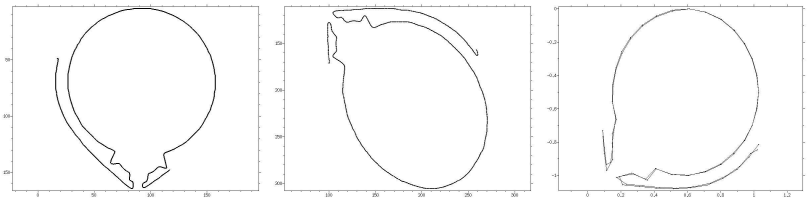
(a) $NFA < 1$



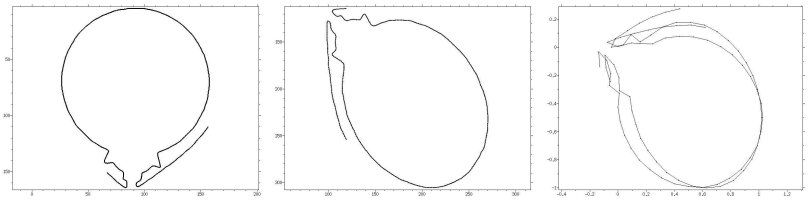
(b) $NFA < 10^{-2}$

Figure 7.2: Affine invariant semi-local recognition: meaningful matches ($NFA < 1$) between shape elements. No false match was detected

Finally, notice that one of the nested boundaries of the symbol does not present any matched shape element, while the other one (which is almost symmetric to it) does. This is due to the fact that, in the scene image, one of these nested boundaries presents a single flat piece leading to an "encodable" shape element (all the others bitangent lines or flat parts do not because the curve is not long enough), while in the other one this flat piece is not detected. Once again, this is not really a problem, since these quasi-convex curves are encoded by the global method presented in Chap. 5, section 5.1.



(a) Best match, $NFA = 5.4\ 10^{-7}$



(b) Worst match, $NFA = 9.6\ 10^{-1}$

Figure 7.3: Affine invariant semi-local recognition: The matches showing the lowest and the largest NFA less than 1. On the right column, both matched codes are superimposed. As expected, the first match is much more accurate

Let us now describe the second part of this experiment, which consists in applying the semi-local similarity invariant recognition method to the same query and scene images that were used in the first part. The similarity invariant method is not expected to do better than the affine invariant one, since the common shape elements in the query and the scene images are related to each other by an affine transform. However, it can be interesting to find out if the semi-local similarity invariant method is able to retrieve some matches. In this second part of the experiment, the same stages than in the previous one were followed, except for the normalization/encoding procedure, where the semi-local similarity invariant encoding method described in Chap. 5 was used. In the query image, 80 shape elements were extracted from its meaningful boundaries, and 127 for the scene image. Notice that the similarity invariant encoding is more redundant than the affine invariant encoding, since more shape elements are extracted in the latter case. The explanation is simple: as pointed out in Chap. 5 (section 5.2), the construction of the affine invariant semi-local frames imposes more constraints on the curve than the one for similarity invariant frames. (These affine semi-local frames are also more global than similarity semi-local frames, what makes them less robust to occlusion). Perfect matches in this second part of the experiment should reach numbers of false alarms as low as $88 \times 127/127^6 = 2.66\,10^{-9}$. Here perfect matches cannot happen, mainly because boundaries are not related to each other by similarity transforms.

All 44 detected meaningful matches between shape elements ($NFA < 1$) for the similarity semi-local invariant recognition method are shown, superimposed, in Figure 7.4. Figure 7.5 displays the matching pieces of level lines and their corresponding shape elements for the largest and the lowest NFA ($2.5\ 10^{-5}$ and $7.1\ 10^{-1}$), as well as another example of matched shape elements.

It can be seen from the superimposed normalized shape elements, that these shape elements are not as close as for the affine encoding. However, just looking at shape elements in Figures 7.5(a) and 7.5(c) is enough to see

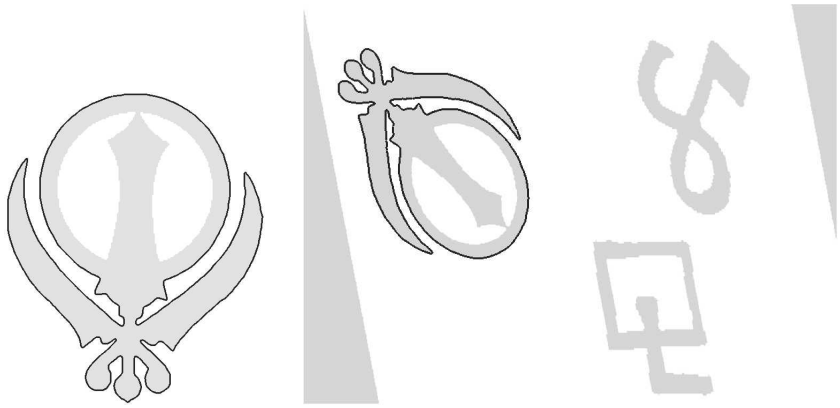
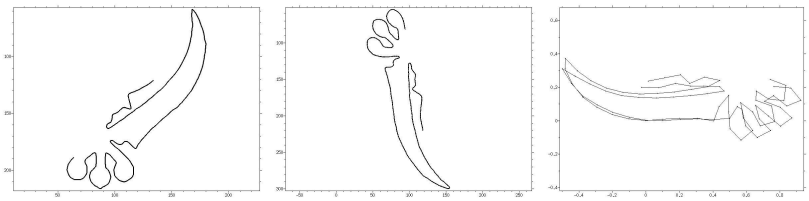
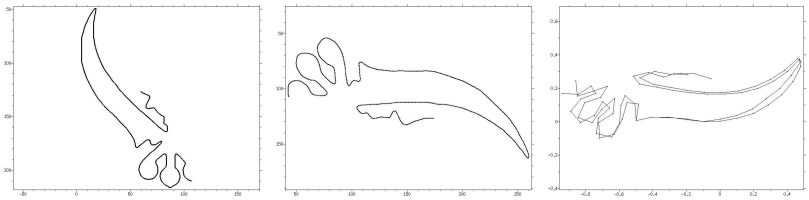


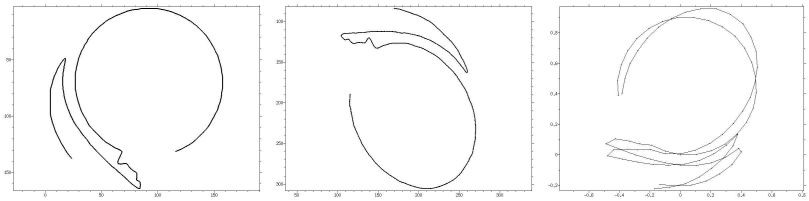
Figure 7.4: Similarity invariant semi-local recognition: Meaningful matches ($NFA < 1$) between shape elements. No false match was detected



(a) Best match, $NFA = 2.5\ 10^{-5}$



(b) Worst match, $NFA = 7.1\ 10^{-1}$



(c) Another example, $NFA = 2.5\ 10^{-4}$

Figure 7.5: Similarity invariant semi-local recognition: the matches showing the lowest and the largest NFA

that, even if the query and the scene images are related by an affine transform presenting non isotropic zooms and a considerable shear, almost the entire shape (except for the nested boundaries, which are "too convex" to be encoded by the semi-local method) can be recognized with a relatively high degree of confidence.

Part of the discussion presented in this section can be summarized in Figure 7.6. The list of meaningful matches is ordered from best (lowest NFA) to worst (largest NFA), and the index i of this sorted list is plotted *versus* $-\log_{10}(NFA_i)$, where NFA_i is the NFA of the i-th best match. Such a function is plotted for the similarity and for the affine matches. The affine semi-local invariant matches reach lower NFA. Notice that in both affine and similarity invariant recognition methods, there are several matches that show small NFA, leading to sure detections of common shapes.

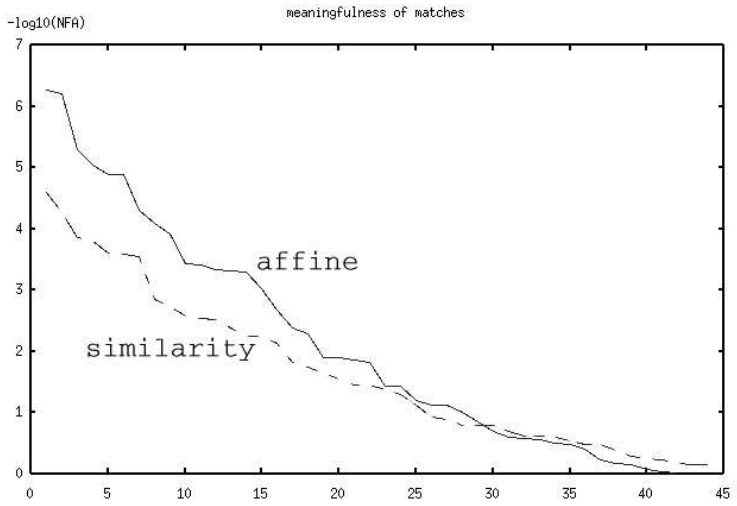


Figure 7.6: NFA of affine and similarity semi-local invariant matches for the toy example. Both lists of meaningful matches are ordered from best (lowest NFA) to worst (largest NFA), and for each list, the index i of the sorted list is plotted *versus* $-\log_{10}(NFA_i)$, where NFA_i is the NFA of the i-th best match

7.1.2 Perspective distortion

It is not surprising that the affine method performs better than the similarity method, when dealing with images related through an affine transform, which do not suffer from occlusion. This second experiment shows that, as expected, the affine method also performs better than the similarity method, when applied to real images related through moderately weak perspective transformations. The two images considered in this experiment (which we call "Hitchcock experiment") are shown in Figure 7.7, with their corresponding level lines. The resolution of these images is 640×480, which is enough to ensure good accuracy in the extracted level lines.

For the affine semi-local invariant method, 1,150 and 853 shape elements were extracted from the query image and from the scene image, respectively. The number of 1-meaningful matches detected was 517. In order to reduce the redundancy of the output, a greedy algorithm eliminates matched shape elements which share a large piece of curve with other shape elements presenting lower NFA. More precisely speaking, if a pair of shape elements $(\mathcal{S}_1, \mathcal{S}'_1)$ is an ε_1-meaningful match, and there exists another pair $(\mathcal{S}_2, \mathcal{S}'_2)$ matching ε_2-meaningfully, with $\varepsilon_2 < \varepsilon_1$, such that $\mathcal{S}_1$ shares at least half of its length with $\mathcal{S}_2$, and if the same property holds for $\mathcal{S}'_1$ and $\mathcal{S}'_2$, then the pair $(\mathcal{S}_1, \mathcal{S}'_1)$ is eliminated from the output list of matches. By this elimination of redundant matches, the list of meaningful matches is drastically reduced from 517 to 16 elements. This also shows how redundant is the encoding. These 16 matched shape elements are shown, superimposed, in Figure 7.8. No false matches were detected, and all matches have their NFA below 0.1. The best match, shown in Figure 7.9, reaches $NFA = 6.5\,10^{-11}$. This value is remarkably low, considering that ideal perfect matches

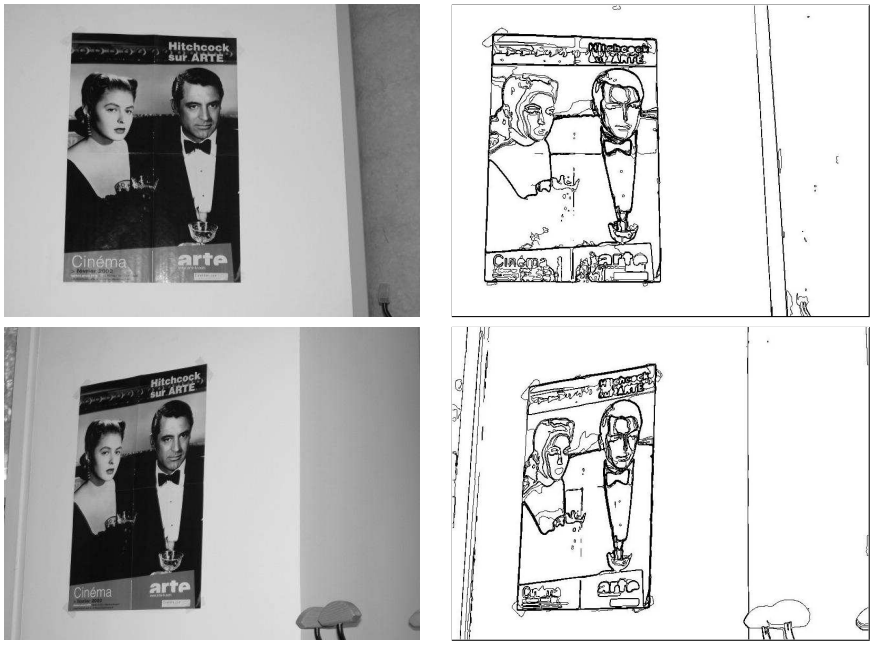


Figure 7.7: Hitchcock experiment: Original images and their corresponding level lines. The image on top is considered as "query" image. In the query image, 307 maximal boundaries were detected, and 266 maximal boundaries were detected in the scene image

in this experiment would have a number of false alarms of $1150 \times 853/853^6 = 2.5\,10^{-12}$ (when the empirical distributions of distances to query shape elements are learned using only the considered scene image, as done here).

Figure 7.8: Affine invariant semi-local recognition method: meaningful matches between shape elements. No false matches were detected, and all detections show an NFA below 0.1. The lowest NFA is $6.5\,10^{-11}$

Figure 7.10 displays the meaningful matches detected using the similarity semi-local invariant recognition method. In this case, 2,033 and 1,463 shape elements were extracted from the query image and from the scene image, respectively. As noticed for the toy example, the similarity method allows to extract more shape elements than the affine method. A total number of 244 meaningful matches ($NFA < 1$) were detected, and 26 matches were left after applying the greedy algorithm. The meaningful matches for the similarity method are shown in Figure 7.10. The lowest NFA reached with the similarity method is $3.8\,10^{-8}$, and corresponds to the shape elements and the normalized shape elements presented in Figure 7.11. Figures 7.10(b) and 7.10(c) present, respectively, the shape elements matching at $\varepsilon < 0.1$, and those for which the NFA is between 0.1 and 1. Notice that none of the 10^{-1}-meaningful matches are false matches, and that the corresponding shape elements are in general much more local than the shape elements matching in Figure 7.10(c).

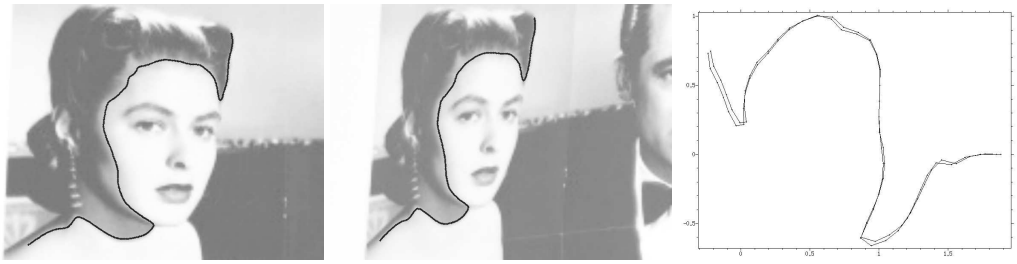


Figure 7.9: Affine invariant semi-local recognition method: The match showing the lowest NFA ($6.5\,10^{-11}$)

Indeed, the more global are the shape elements, the less accurate is the similarity approximation of the underlying transformation, which is in fact a projective transform. Two false matches, for which the NFA is larger than 0.1, can be seen in Figure 7.10(c). Figure 7.12 shows the shape elements of these false matches, as well as the superimposed normalized shape elements represented in the normalized frame.

We end up the discussion on the "Hitchcock experiment" with a comparison between the NFA of the meaningful matches for the affine and the similarity semi-local invariant methods. Such a comparison is illustrated in Figure 7.13. The principle is the same as for the toy example from section 7.1.1. The list of meaningful matches is ordered from best (lowest NFA) to worst (largest NFA), and the index i of this sorted list is plotted *versus* $-\log_{10}(NFA_i)$, where NFA_i is the NFA if the i-th best match. Such a function is plotted for the similarity and for the affine matches. The affine semi-local invariant matches reach lower NFA. Notice that in both affine and similarity invariant recognition methods, there are several matches that show small NFA, leading to sure detections of common shapes.

7.1.3 A more difficult problem

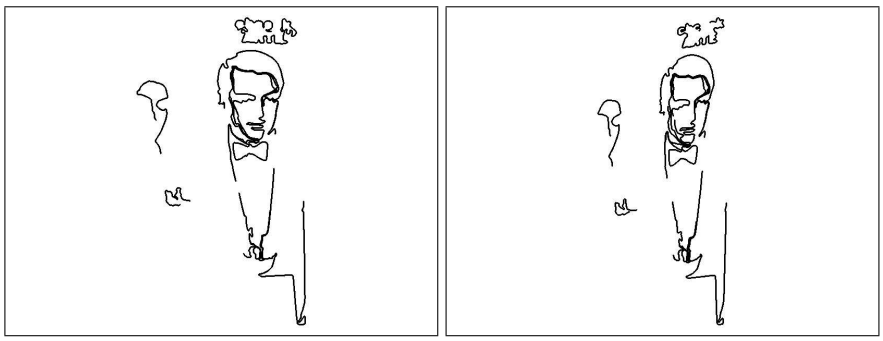
Both in the toy example and the in "Hitchcock experiment", query and scene images represented different views of the same planar "objects" or elements. Corresponding shapes were accurately described by the meaningful boundaries, leading to the detection of several matching shape elements, with high detection confidence. In this subsection, a more difficult example is considered. It consists in finding common shape elements between the pair of images in Figure 7.14. Although at first sight these two different posters of the movie *Casablanca* are very similar, they present many differences that considerably affect the topographic map, and consequently the set of maximal meaningful boundaries. For instance, the actors' faces in the query image (the one on top in Figure 7.14) come from a snapshot, while the scene image is a drawing.

For this example, only the similarity semi-local invariant method is considered. The number of shape elements that were extracted from the query and the scene images were 3,540 and 8,554, respectively. Figure 7.15 shows the 1-meaningful matches (*i.e.* matches for which $NFA < 1$) in the top row, and the 10^{-1}-meaningful matches in the bottom row. The number of detected 1-meaningful matches was 211, which was reduced to 17 after applying the greedy algorithm. It seems that the majority of the relevant shape information that both images have in common has been detected. No meaningful match was found for characters 'Casab', which are indeed quite different (up to similarity invariance) in both images.

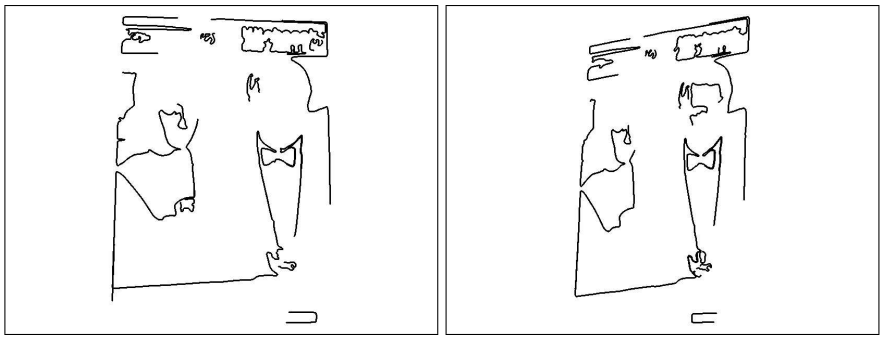
Figure 7.16 shows the shape elements corresponding to the most meaningful match, for which $NFA = 1.1\,10^{-6}$. Such a low NFA is a consequence of the fact that this query shape element is so singular, that it is almost impossible that just by chance another shape element lies so close to it. At this point, it is worth making the following remark, which is obvious from the definition of the NFA given in Chap. 6. Suppose that two query shape elements $\mathcal{S}_1$ and $\mathcal{S}_2$ and two scene shape elements $\mathcal{S}'_1$ and $\mathcal{S}'_2$ are given, such that $d(\mathcal{S}_1, \mathcal{S}'_1) = d(\mathcal{S}_2, \mathcal{S}'_2) = \delta$. Then, if

$$\#\{\mathcal{S}' \in \mathcal{B} \text{ s.t. } d(\mathcal{S}_1, \mathcal{S}') \leqslant \delta\} < \#\{\mathcal{S}' \in \mathcal{B} \text{ s.t. } d(\mathcal{S}_2, \mathcal{S}') \leqslant \delta\},$$

(a) All 26 matches having an NFA below 1



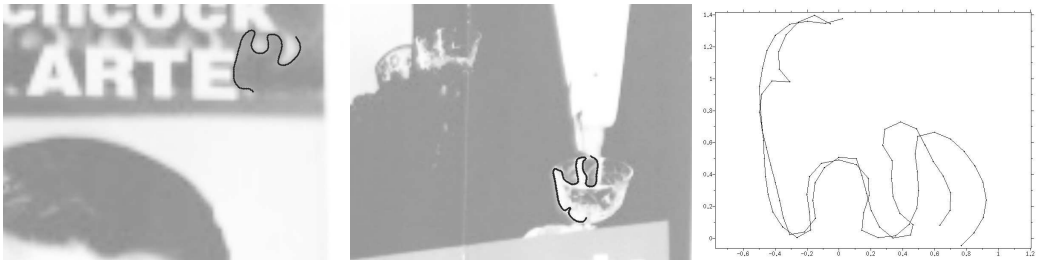
(b) 12 matches show an NFA below 0.1



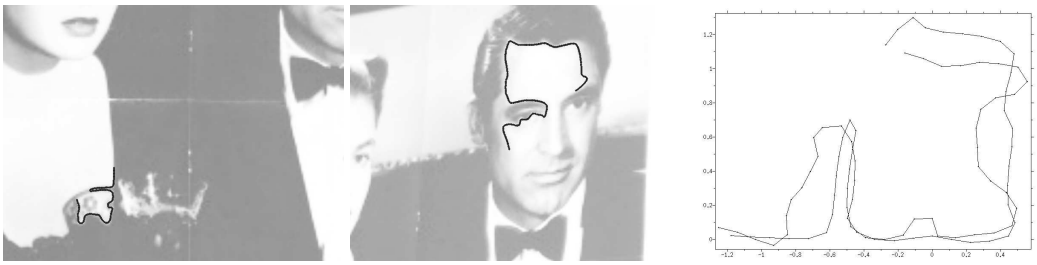
(c) 14 matches show an NFA between 0.1 and 1

Figure 7.10: Similarity invariant semi-local recognition method: Meaningful matches between shape elements. Among the 26 matches having an NFA below 1, 12 are 10^{-1}-meaningful. Two false matches can be seen in (c); their NFA is above 0.1

Figure 7.11: Similarity invariant semi-local recognition method: The match showing the lowest NFA ($3.8\,10^{-8}$)



(a) False match, $NFA = 0.64$



(b) False match, $NFA = 0.68$

Figure 7.12: Similarity semi-local invariant method: The two false matches. Their NFA are larger than 0.1

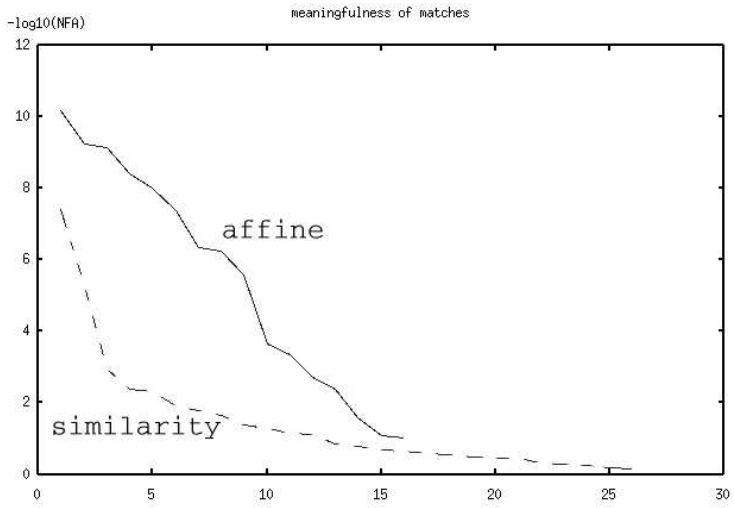
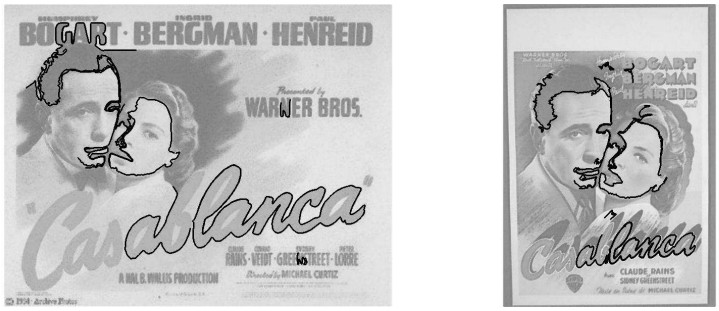
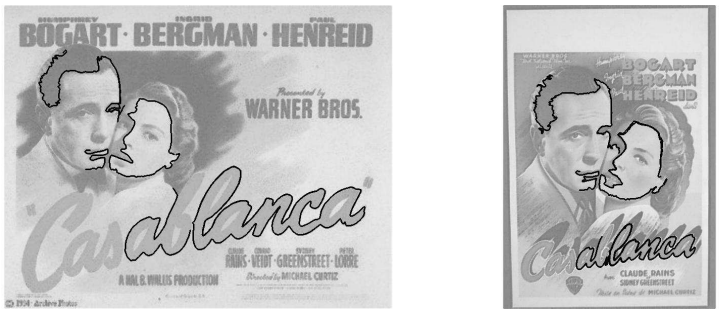


Figure 7.13: Hitchcock experiment: NFA of affine and similarity semi-local invariant matches. Both lists of meaningful matches are ordered from best (lowest NFA) to worst (largest NFA), and for each list, the index i of the sorted list is plotted *versus* $-\log_{10}(NFA_i)$, where NFA_i is the NFA of the i-th best match

Figure 7.14: Casablanca experiment: Original images (on the left) and level lines (on the right). The image on top was considered as query image



(a) $NFA < 1$



(b) $NFA < 0.1$

Figure 7.15: Casablanca experiment: Meaningful matches between similarity invariant shape elements. Top: $NFA < 1$. Bottom: $NFA < 10^{-1}$, no false match can be seen

it follows that $NFA(\mathcal{S}_1, \mathcal{S}'_1) < NFA(\mathcal{S}_2, \mathcal{S}'_2)$. Hence, for a given distance d, the "rarer" a query shape element $\mathcal{S}$ is (with respect to $\mathcal{B}$), the lower is $NFA(\mathcal{S}, d)$. This makes sense, since a rare shape element is more discriminatory than a banal one.

Figure 7.16: The match with the lowest NFA. The query shape element (left image) matches the database shape (right image). $NFA = 1.1 \cdot 10^{-6}$

Figure 7.17 shows all the false matches detected at $NFA < 1$. They all have an NFA between 0.1 and 1.

Finally, notice that all matches which semantically correspond to the same shape elements (the "correct" matches) show NFAs below 0.1.

Figure 7.17: The five false matches for $NFA < 1$, with their normalized shape elements. The left most and middle images correspond to the query and the scene shape elements, respectively. The right most image shows their normalized shape elements, superimposed. All false matches show NFAs between 0.1 and 1

7.1.4 Slightly meaningful matches between unrelated images

The experiment presented in this subsection consists in looking for common shape elements between two unrelated images. Two examples are considered. "Query" and "scene" images for the first experiment are shown in Figure 7.18. All the matches for which NFA is below 1 are displayed, superimposed to the original images. 4,731 and 4,946 shape elements were extracted from query and scene images, respectively. Among all $4731 \times 4946 \approx 23\,10^6$ pairs of query-scene shape elements, only 6 matches having $NFA < 1$ were detected. Their NFAs range from 0.21 to 0.97. The matched shape elements, as well as their corresponding normalized shape elements, are shown in Figure 7.19. Numbers $1)$, $4)$ and $5)$, are "simple" (they are relatively short and do not present many oscillations), and match with pretty small distances. However, because of their "banality", they do not show lower NFAs. Matches number $2)$ and $6)$, while locally different, are quite similar at a coarse scale, as can be seen from their superimposed normalized shape elements. For such long shape elements, a representation using 45 points may not be accurate enough; a finer sampling would probably have led to larger NFAs for that kind of matches.

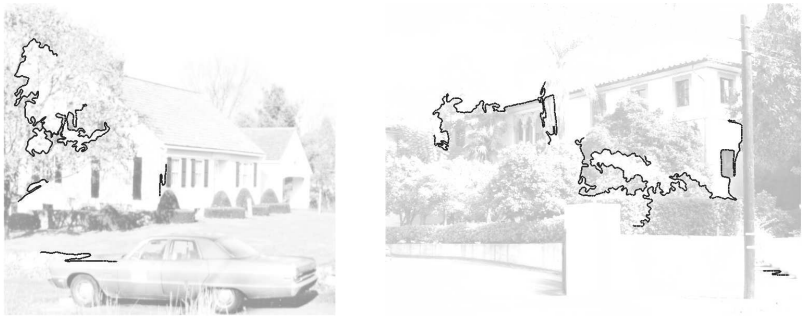


Figure 7.18: Left: Query image; 4731 shape elements were extracted from this image. Right: Scene image; the number of shape elements extracted from it was 4946. Among the $23\,10^6$ pairs of query-scene shape elements, only six match with $NFA < 1$. The NFA of these matches range from 0.21 to 0.97

The second example of common shape elements between two unrelated images involves the images in Figure 7.20. The 22 shape elements extracted from the query image are searched in the 546 shape elements from the scene image on the left. The two shape elements that match with $NFA < 1$ are shown, superimposed to the original images. Unlike the previous example, here these matches show NFAs lower than 0.1. The matched shape elements and their normalized shape elements are shown in Figure 7.21. Notice that, according to what was presented in Chap. 6, matches showing NFAs lower than 0.1 are not supposed to happen "by chance" (as matches between shape elements extracted from random level lines), and some common cause should be behind such an unexpected coincidence. This is what happens here. Indeed, many shapes in images derive from natural or man-made objects having a common structure. For instance, many objects are built of parallel or equal-length parts.

7.1.5 Blur introduced by long distances to the camera

In this subsection we present the last experiment dealing with the semi-local invariant method for shape recognition. This experiment just aims at illustrating how the meaningful boundaries of small objects are affected by the blur introduced when objects are far from the camera, and how this problem can be solved by representing the query image at multiple scales.

The query and scene images for this example are shown in Figure 7.22, with their corresponding maximal meaningful boundaries. Images are displayed at the same scale. Figure 7.23 illustrates a detail of the maximal meaningful boundaries of the scene image, corresponding to the region of interest for this experiment. Compare now these boundaries with the ones extracted from the query image, in Figure 7.22 (on top right). The

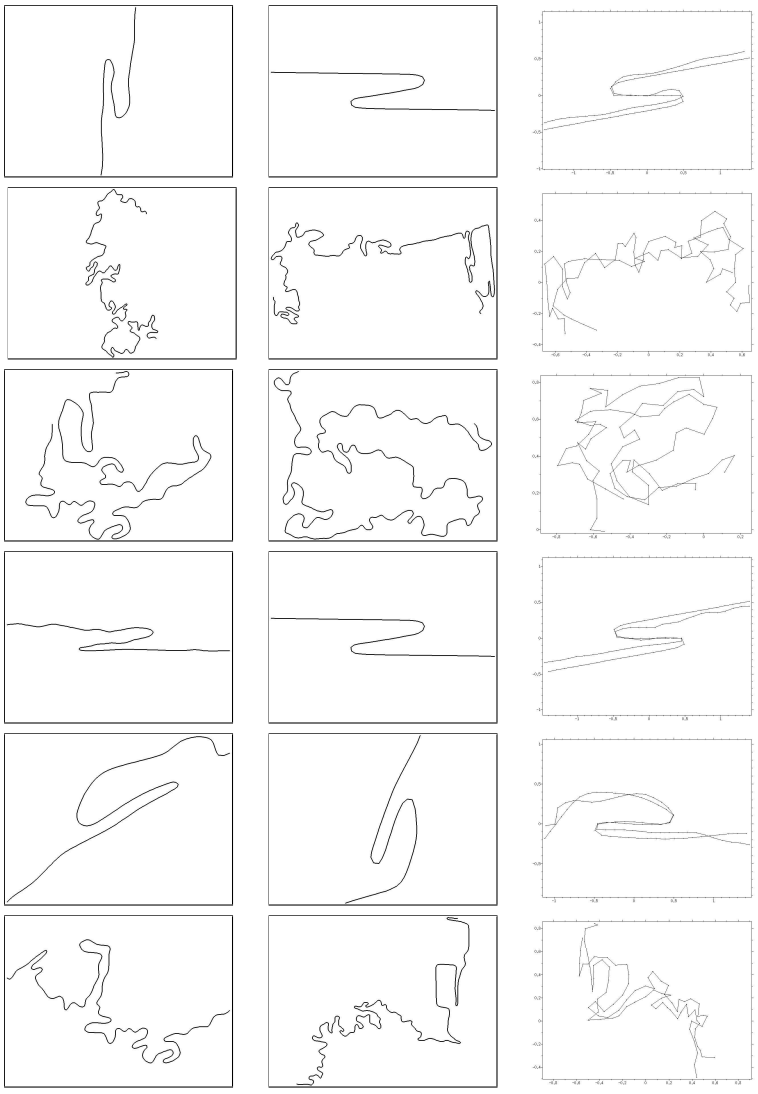


Figure 7.19: The six false matches detected for $NFA < 1$, with their normalized shape elements. The left most and middle images correspond to the query and the scene shape elements, respectively. The right most image shows their normalized shape elements, superimposed. All false matches show NFAs between 0.1 and 1

Figure 7.20: Puma experiment. Left: Query image, from which 22 shape elements were extracted. Right: Scene image; 546 shape elements were extracted from it. The two matches detected at $NFA < 1$ are superimposed to the original images

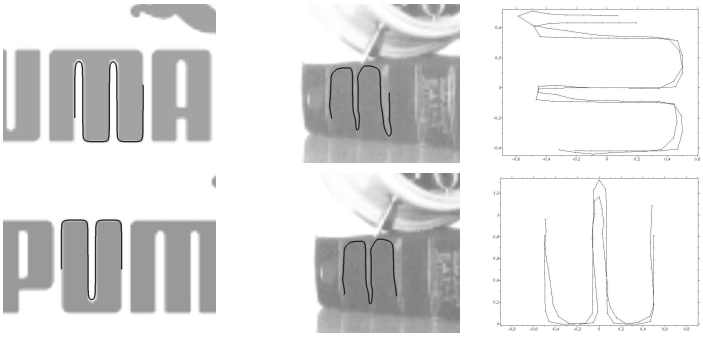


Figure 7.21: Puma experiment: The two matches detected for $NFA < 1$, with their normalized shape elements. The left most and middle images correspond to the query and the scene shape elements, respectively. The right most image shows their normalized shape elements, superimposed. Such a conspicuous coincidence admits a better explanation than randomness: Many shapes in images derive from natural or man-made objects having a common structure. For instance, many objects are built of parallel or equal-length parts

characters in the scene image have been almost completely destroyed, and not many similar shape elements can be observed.

Figure 7.22: Top row: Query image and its maximal meaningful boundaries; 312 shape elements were extracted from this image. Bottom row: Scene image and corresponding maximal meaningful boundaries. 1,859 shape elements were extracted from it

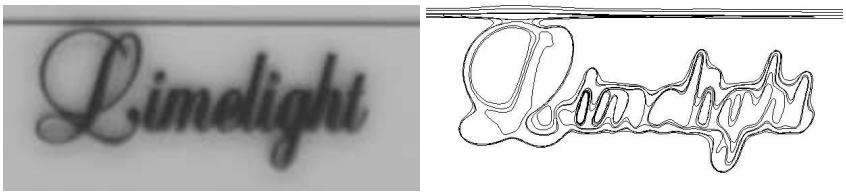


Figure 7.23: detail of the maximal meaningful boundaries of the scene image, corresponding to the region of interest for this experiment. The boundaries of characters have been very degraded by the blur and the smoothing

Figure 7.24 shows, on the top row, the original image and two image reductions, by factors 4 and 8. The bottom row presents their corresponding maximal meaningful boundaries (followed by an affine shortening at scale $T = 0.5$, see Chap. 4, Section 4.3). Image reductions were performed using a prolate kernel.

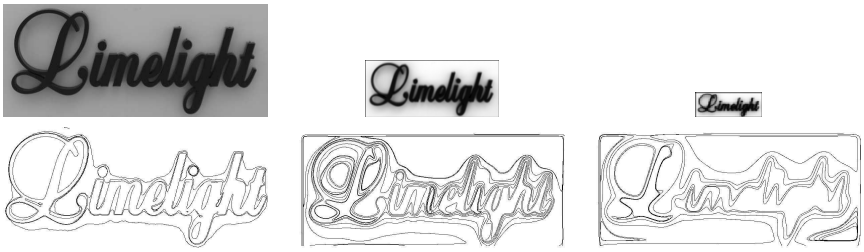
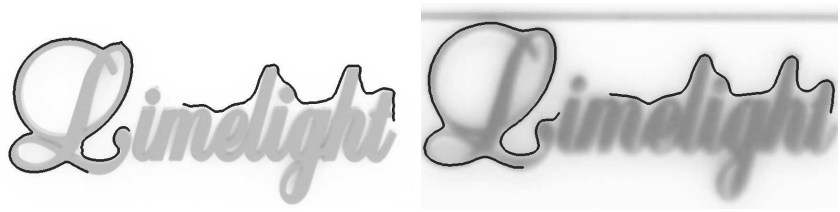
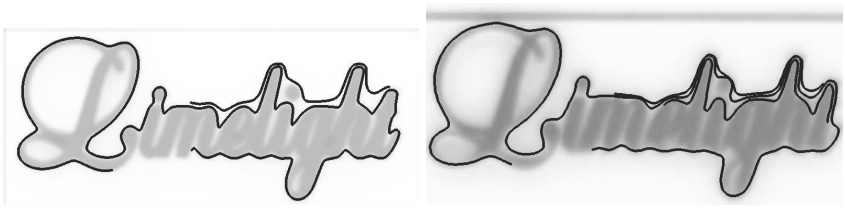


Figure 7.24: Original query image and two image reductions. Left column: Original image and corresponding maximal meaningful boundaries. Middle: Image reduction by a factor 4 (324 shape elements were extracted from this image). Right: Image reduction by a factor 8 (73 shape elements were extracted from this image). Reductions were performed using a prolate kernel

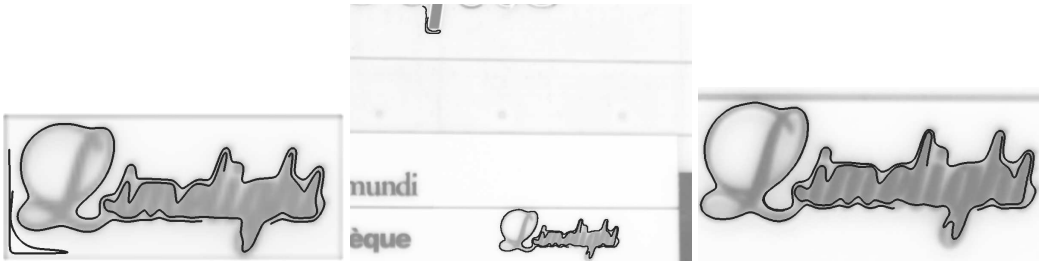
Figure 7.25 shows the detected matches at $NFA < 1$, for each query image (the original image and the two reductions) with the scene image. When the shape elements of the original query image are searched, only two matches having $NFA < 1$ are found (Figure 7.25(a)). Both matches are correct, and their NFAs are $8.8\,10^{-6}$ and $1.9\,10^{-4}$. Using as query image the image reduced by a factor 4, more meaningful matches are found, and the best one reaches an NFA of $2.3\,10^{-10}$. In this case all matches are correct (Figure 7.25(b)). Finally, using the query image reduced by a factor 8, even more meaningful matches are detected, reaching lower NFAs. In this case, the NFA of correct matches ranges from $2.1\,10^{-3}$ to $3.6\,10^{-12}$. A false match of $NFA = 7.6\,10^{-1}$ was detected, but it corresponds to an artifact (a border effect) of the image reduction, as can be seen in Figure 7.25(c).



(a) Using the original query image. 2 matches have their NFA below 1 ($8.8\,10^{-6}$ and $1.9\,10^{-4}$)



(b) Using an image reduction by 4 of the query image: 4 meaningful matches, with NFAs equal to $2.3\,10^{-10}$, $1.3\,10^{-6}$, $1.5\,10^{-5}$ and $5.7\,10^{-1}$



(c) Using an image reduction by 8 of the query image: 5 meaningful matches, at $NFAs$ $3.6\,10^{-12}$ $7.4\,10^{-5}$, $4.6\,10^{-4}$, $2.1\,10^{-3}$ and $7.6\,10^{-1}$. The last one corresponds to a false match, but was introduced by an artifact in the image reduction procedure

Figure 7.25: Shape elements matched with the scene images, using three different scales of a query image. The number of meaningful matches, as we their meaningfulness, increases when we consider image reductions. These image reductions simulate the effect of distance to the camera

7.2 Recognition is relative to the context

This section illustrates a property of the distance threshold derived from the number of false alarms: the distance threshold for recognition should depend on the "rareness" or on the "banality" of the target shape element relative to a database of shape elements, and therefore on the context.

The aim of the experiments presented here is to validate the previous claim. Four shape elements extracted from the character 'm' (Figure 7.26) are sought for in 14 scanned pages, by using the semi-local similarity invariant recognition method.

Two experiments were led: In the first one the database that was used to learn probabilities consisted of these 14 scanned pages (79,376 shape elements), whereas in the second one the database was made of shape elements extracted from 21 'natural' images (98,857 shape elements).

Figure 7.27 shows the shape elements from one of the 14 scanned pages that matched with the target shape elements, when probabilities are estimated over the scanned text database (notice that all 'm' are recognized).

Figure 7.28 shows the recognition result when the scanned text database is replaced by a "natural image" database. We can see that the recognition thresholds are more permissive in the second case (Figure 7.29). This result is fully coherent with the theory: in the first case, the focus is put on recognition of shape elements that share some common structure with 'm' *among other characters*, that is to say other 'm', whereas in the second case, the focus is put in recognizing shape elements that share a common structure with 'm' *relative to a large universe of shape elements extracted from natural images*, that is to say other 'similar' characters (this explains why italic 'm' and other less similar shapes are retrieved).

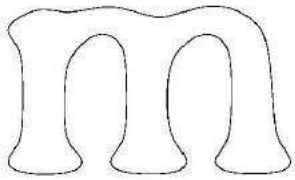
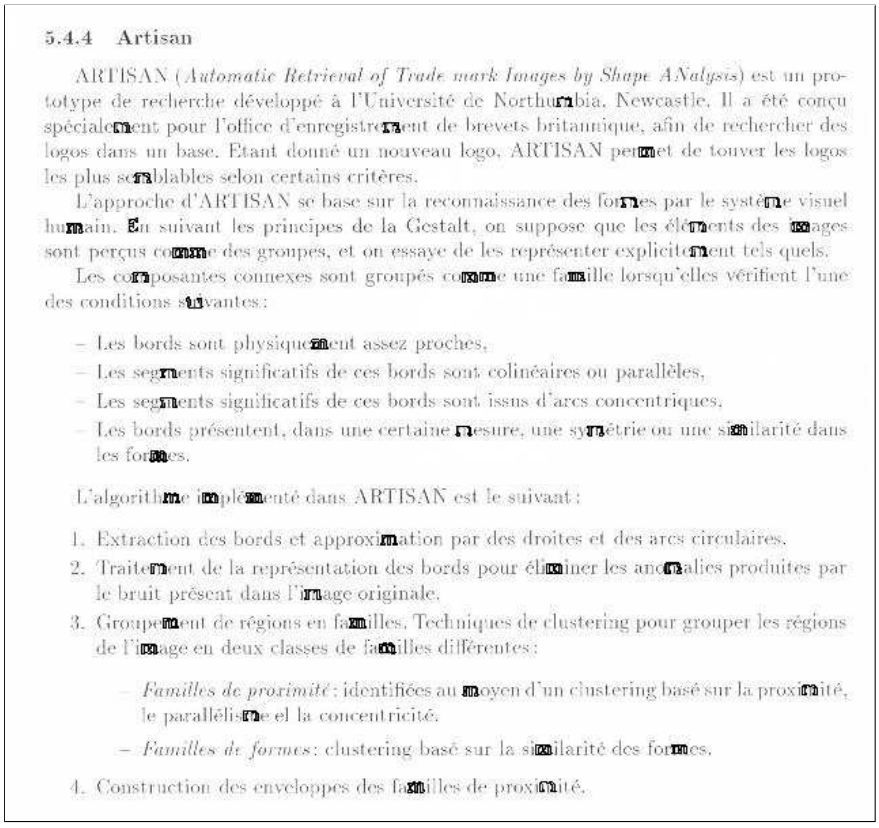


Figure 7.26: Characters - the query curve



5.4.4 Artisan

ARTISAN (*Automatic Retrieval of Trade mark Images by Shape ANalysis*) est un prototype de recherche développé à l'Université de Northumbia, Newcastle. Il a été conçu spécialement pour l'office d'enregistrement de brevets britannique, afin de rechercher des logos dans un base. Etant donné un nouveau logo, ARTISAN permet de touver les logos les plus semblables selon certains critères.

L'approche d'ARTISAN se base sur la reconnaissance des formes par le système visuel humain. En suivant les principes de la Gestalt, on suppose que les éléments des images sont perçus comme des groupes, et on essaye de les représenter explicitement tels quels.

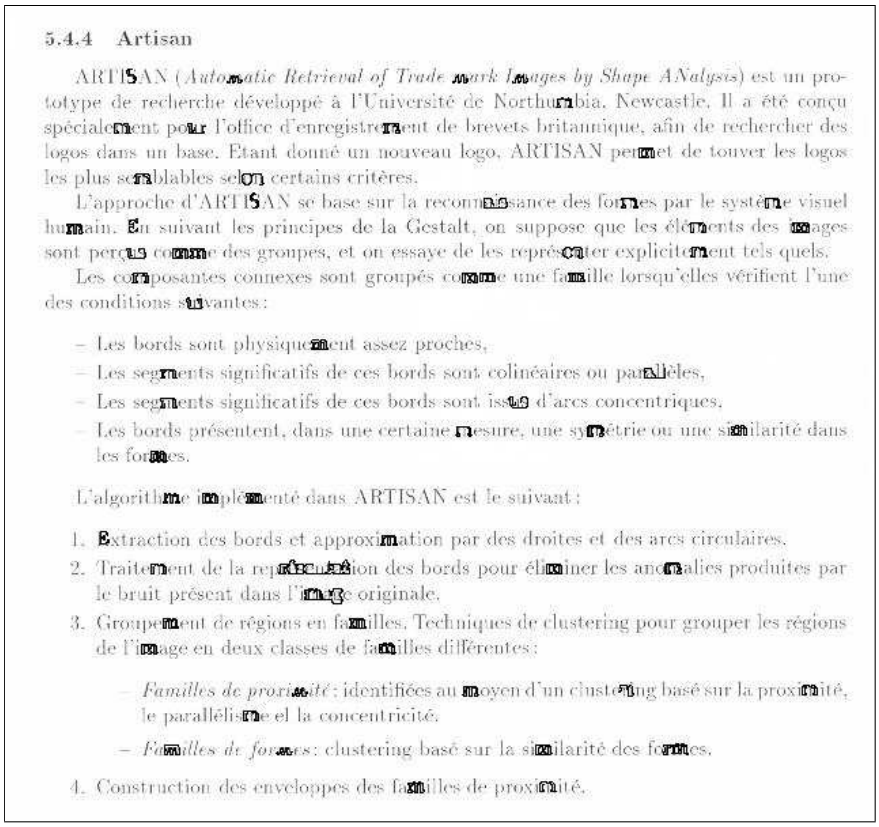
Les composantes connexes sont groupés comme une famille lorsqu'elles vérifient l'une des conditions suivantes:

- Les bords sont physiquement assez proches,
- Les segments significatifs de ces bords sont colinéaires ou parallèles,
- Les segments significatifs de ces bords sont issus d'arcs concentriques.
- Les bords présentent, dans une certaine mesure, une symétrie ou une similarité dans les formes.

L'algorithme implémenté dans ARTISAN est le suivant:

1. Extraction des bords et approximation par des droites et des arcs circulaires.
2. Traitement de la représentation des bords pour éliminer les anomalies produites par le bruit présent dans l'image originale.
3. Groupement de régions en familles. Techniques de clustering pour grouper les régions de l'image en deux classes de familles différentes:
   - *Familles de proximité*: identifiées au moyen d'un clustering basé sur la proximité, le parallélisme el la concentricité.
   - *Familles de formes*: clustering basé sur la similarité des formes.
4. Construction des enveloppes des familles de proximité.

Figure 7.27: Characters - Recognition when probabilities are estimated over the database of scanned text pages. A total number of 111 matches were detected. All 'm''s having the same font type that the query where retrieved. Only two matches with shape elements which do not belong to an 'm' were found, but these curves are really close up to similarity invariance



5.4.4 Artisan

ARTISAN (*Automatic Retrieval of Trade mark Images by Shape ANalysis*) est un prototype de recherche développé à l'Université de Northumbia, Newcastle. Il a été conçu spécialement pour l'office d'enregistrement de brevets britannique, afin de rechercher des logos dans un base. Etant donné un nouveau logo, ARTISAN permet de touver les logos les plus semblables selon certains critères.

L'approche d'ARTISAN se base sur la reconnaissance des formes par le système visuel humain. En suivant les principes de la Gestalt, on suppose que les éléments des images sont perçus comme des groupes, et on essaye de les représenter explicitement tels quels.

Les composantes connexes sont groupés comme une famille lorsqu'elles vérifient l'une des conditions suivantes:

- Les bords sont physiquement assez proches,
- Les segments significatifs de ces bords sont colinéaires ou parallèles,
- Les segments significatifs de ces bords sont issus d'arcs concentriques,
- Les bords présentent, dans une certaine mesure, une symétrie ou une similarité dans les formes.

L'algorithme implémenté dans ARTISAN est le suivaat:

1. Extraction des bords et approximation par des droites et des arcs circulaires.
2. Traitement de la représentation des bords pour éliminer les anomalies produites par le bruit présent dans l'image originale.
3. Groupement de régions en familles. Techniques de clustering pour grouper les régions de l'image en deux classes de familles différentes:
   - *Familles de proximité*: identifiées au moyen d'un clustering basé sur la proximité, le parallélisme el la concentricité.
   - *Familles de formes*: clustering basé sur la similarité des formes.
4. Construction des enveloppes des familles de proximité.

Figure 7.28: Characters - Recognition when probabilities are estimated over the database extracted from natural images. 154 matches were detected. The corresponding distance thresholds obtained in this case were indeed larger than those in Fig. 7.27

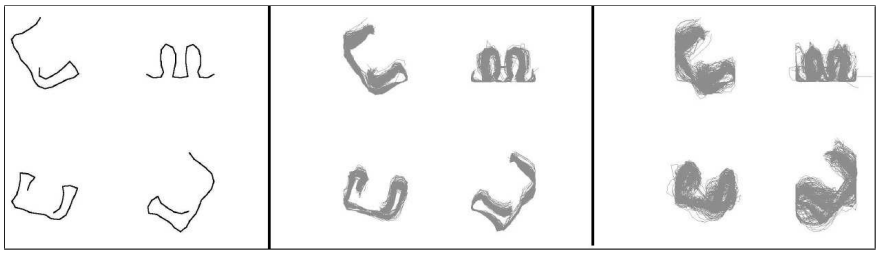


Figure 7.29: Characters - Superimposition of the matched normalized shape elements. Left: the four target shape elements. Middle: all shape elements from the scanned text that match the corresponding target shape element, superimposed; probabilities were estimated using the 14 scanned pages. Right: superimposed matched shape elements when probabilities were estimated over the database of natural images. Notice that the matching threshold is larger in the latter case

7.3 Global meaningful matches

In Chap. 5 normalizations invariant up to similarity or affine transforms were presented. This section shows several experiments on global matching of shapes that validate the normalization and the distance threshold derived from the number of false alarms (see chapter 6). Before going to the experiments, let us recall that a curve leads to as many descriptions (or "codes") as bitangent or flat parts are present in the curve.

7.3.1 Global affine invariant recognition: toy example

This first experiment illustrates the global recognition method with a simple example. The pair of images considered here involves the two images presented in Figure 7.30, where the global meaningful matches have been superimposed. The image on the left was considered as query image, and the one on the right contains an affine distorted version of the query. Global shape elements were extracted by means of the global affine invariant normalization method (Chap. 5, Section 5.1.2), after extracting the meaningful boundaries and smoothing them with the affine curve shortening.

The detection of meaningful matches between all global shape elements extracted from both images was performed using the detection method presented in Chap. 6. A single false match was detected, with an NFA of 0.53. Matches between four different pairs of curves were detected. For each of these pairs, the best match (recall that between two curves, several matches between global shape elements may exist) is shown in Figure 7.31, with their corresponding NFA.

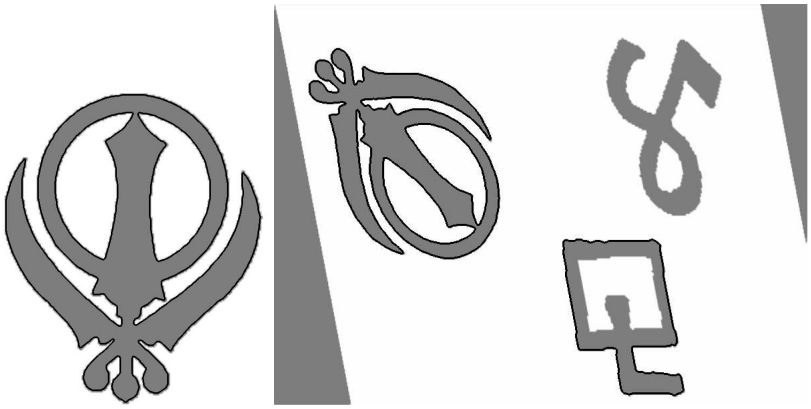


Figure 7.30: Toy example: Original images with global meaningful matches, superimposed. The image on the right contains an affine distorted version of the symbol in the left image

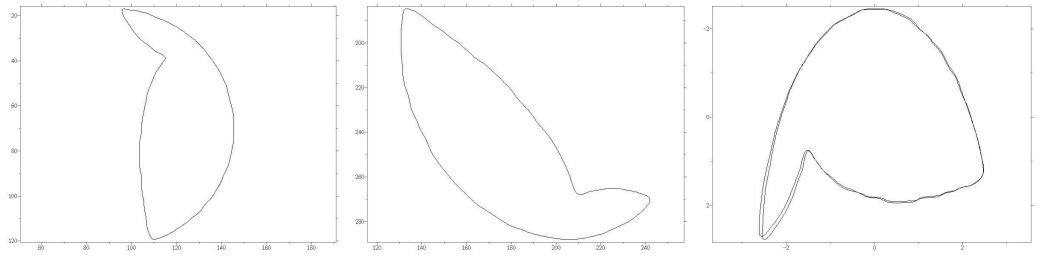
7.3.2 Comparing similarity and affine invariant global recognition methods

In this experiment we compare the performance of the affine and the similarity global recognition methods, on two images whose maximal meaningful boundaries are shown in Figure 7.32. The boundaries on the left were considered as query. The shapes in the scene images are distorted not only by projective transforms but also from projection on a cylinder (the bottle).

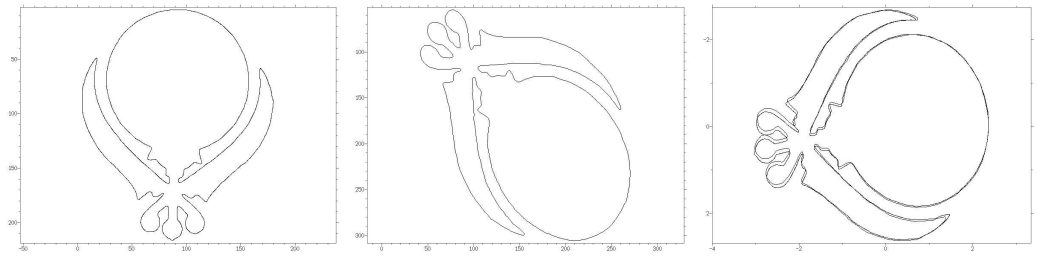
The example considered here consists in finding the character 'n' from the query image in the scene image.

Figure 7.33 shows the detected 1-meaningful matches with global shape elements extracted from the 'n' in the query image, for the similarity invariant method. The query 'n' was represented with 10 global shape elements. 36 matches were found in the scene image. The lowest NFA was 10^{-11}. Some false matches can be seen, but they all show NFAs between 0.7 and 1.

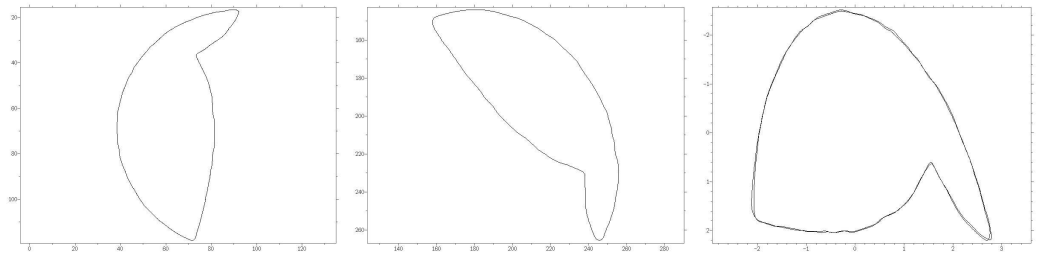
Figure 7.34 shows the matched global shape elements when considering the global affine method. The query 'n' is still represented with 10 shape elements, since it is the same 'n' that was considered for the global



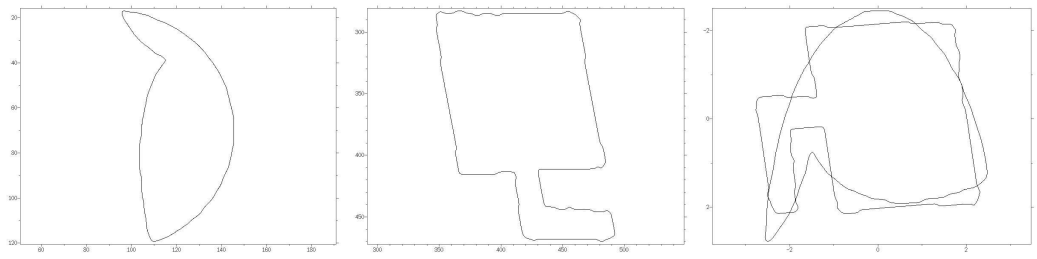
(a) $NFA = 3.7\ 10^{-9}$



(b) $NFA = 4.1\ 10^{-8}$



(c) $NFA = 3.6\ 10^{-7}$



(d) $NFA = 5.3\ 10^{-1}$

Figure 7.31: Affine invariant global recognition: all pairs of curves showing matches with $NFA < 1$. On the right column, the shape elements extracted from the curves that match with the lowest NFA are displayed, superimposed

Figure 7.32: Evian: maximal meaningful boundaries. Left: query. Right: scene

similarity matching (and there is always one global shape element extracted for each bitangent line or flat part of the curve). 35 matches showed an NFA below 1. The matches that actually correspond to the 'n' on the bottle, show NFAs that range from 10^{-15} to 10^{-8}. The NFA of matches which do not correspond to global shape elements in the 'n' on the bottle, are between 10^{-3} and 1. However, some "false matches" are not really "false" but "casual" matches, since they correspond to other characters 'n' or 'u' that appear on the bottle ("Minérale" and "Naturelle").

Figure 7.33: Evian: global similarity invariant matching. All 1-meaningful matches with character 'n' from the query image. The query 'n' is represented with 10 global shape elements, that match with 36 global shape elements from the scene image. The lowest NFA is 10^{-11}. False detections show NFAs between 0.7 and 1

Figure 7.35 shows, for both methods, the matches showing the lowest NFA. The top row shows the normalized global shape element for the global similarity invariant method, and the bottom row shows the normalized global shape element for the affine method. Notice that the pair of affine normalized shape elements are much closer to one another than the pair of similarity normalized shape elements. It seems then reasonable that the NFA reached with the global affine invariant method (10^{-15}) is lower than the one reached with the similarity method (10^{-8}).

Figure 7.34: Evian: affine invariant global matching. Meaningful matches with character 'n' from the query image, represented with 10 global shape elements. Left: 1-meaningful matches, 35 matches. False matches show an NFA between 10^{-3} and 1, but some of them are not really "false" but "casual" matches, since they correspond to other characters 'n' and 'u' which are present in the scene. Good matches show NFA ranging from 10^{-15} to 10^{-8}. Right image: the 23 meaningful matches showing NFAs below 10^{-2}

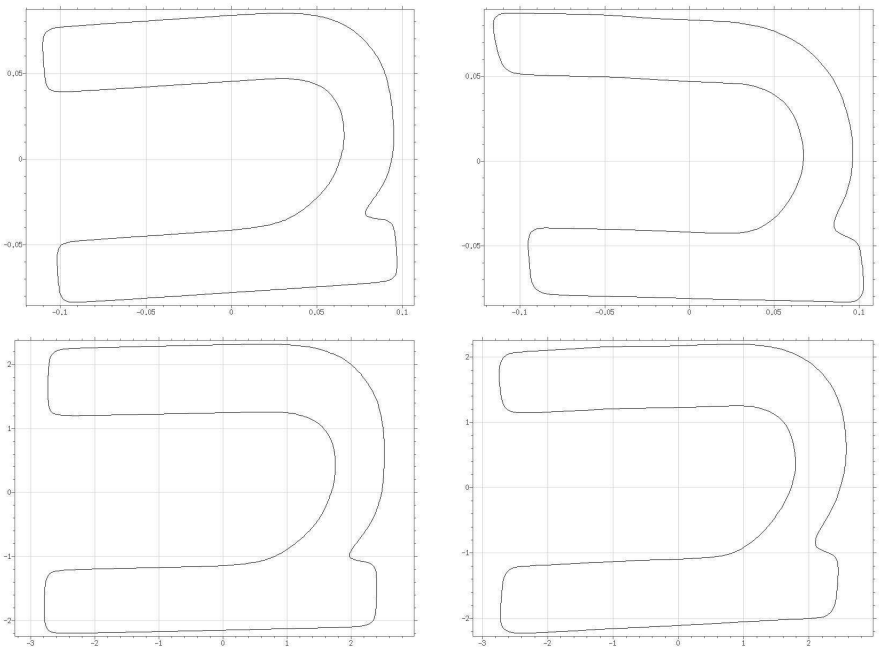


Figure 7.35: Evian experiment: matches for the 'n' showing the lowest NFA for the global similarity (top row) and affine (bottom row) invariant recognition methods. In each of the rows, the curve on the left is the normalized global shape element extracted from the query 'n', and the one on the right is the corresponding normalized global shape element extracted from the scene image. The NFA for the similarity method was 10^{-11}, and for the affine method was 10^{-15}. In spite of the projection on the bottle, the normalized shapes elements are very alike

In Figure 7.36, the false match that show the lowest NFA are presented, for both methods. The top row shows the normalized global shape element for the global similarity invariant method, and the bottom row shows the normalized global shape element for the affine method. The NFA for the similarity invariant match was 0.7, and for the affine method was $4.0\,10^{-3}$.

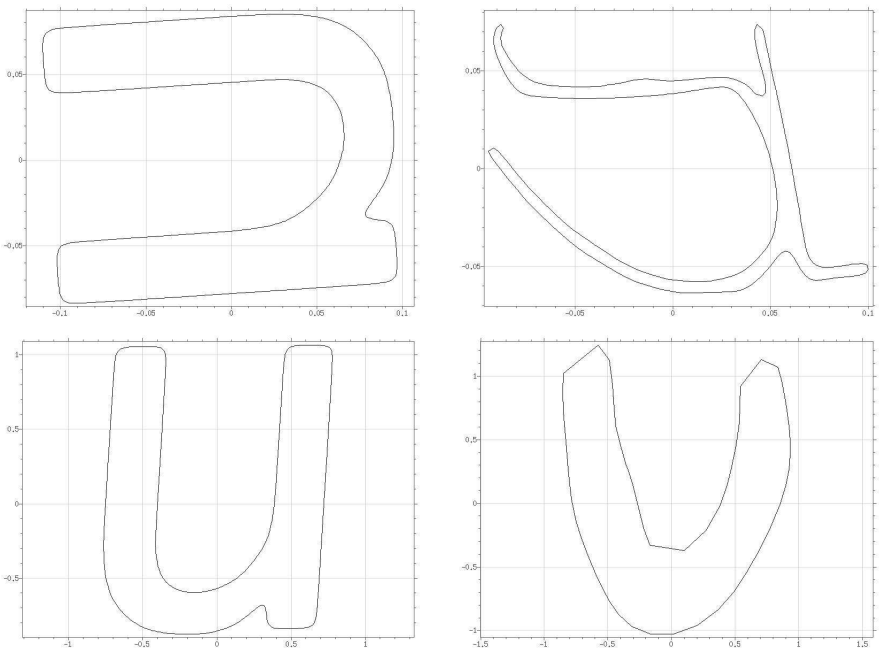


Figure 7.36: Evian experiment: the false matches for the 'n' that shows the lowest NFA, for the global similarity (top row) and affine (bottom row) invariant recognition methods. In each of the rows, the curve on the left is the normalized global shape element extracted from the query 'n', and the one on the right is the corresponding normalized global shape element extracted from the scene image. The NFA for the similarity invariant match was 0.7, and for the affine method was $4.0\,10^{-3}$ (it can be seen in the handwritings on the top of the right image from Figure 7.34)

7.3.3 Global matches of non-locally encoded shapes elements

The main drawback of global shape matching is its sensitivity to occlusions, whereas local matching is especially designed to deal with them. Nevertheless, the semi-local encoding presented in Chap. 5 is unable to encode curves which are convex or "quasi-convex" (that is to say curves for which the length after normalization is not large enough to be encoded). While in general (as will be shown with some experiments) these "quasi-convex" boundaries are not very discriminatory because they are not rich in details, some of them may provide useful information we would not like to miss. Indeed, individually speaking, each match may not be very meaningful, while the conjunction of several of them may be very significant. Thus, the global and semi-local methods must work together: the non-locally encoded boundaries are globally encoded, and then globally compared.

First example: a book cover

Figure 7.37 shows two different views of a book cover, and its corresponding maximal meaningful boundaries. The "query" image (on top) consists of a partial view.

The two images are related by a strong perspective deformation. Perspective transforms can be locally approximated by affine transforms. Indeed many boundaries in images are quite local. It is therefore sound

Figure 7.37: Book cover. Top row: "query" image, and its corresponding 208 maximal meaningful boundaries. Bottom row: "scene" image, and its 1185 maximal meaningful boundaries

to try to find correspondences between the considered pair of images using the semi-local or the global affine invariant recognition methods.

Figure 7.38 shows the 1-meaningful matches between shape elements detected by the semi-local affine recognition method. Among the 16 matches that were found, a single false match having an NFA equal to 0.6 was detected (it can be seen on the right part of the scene image), and the lowest NFA was 10^{-10}.

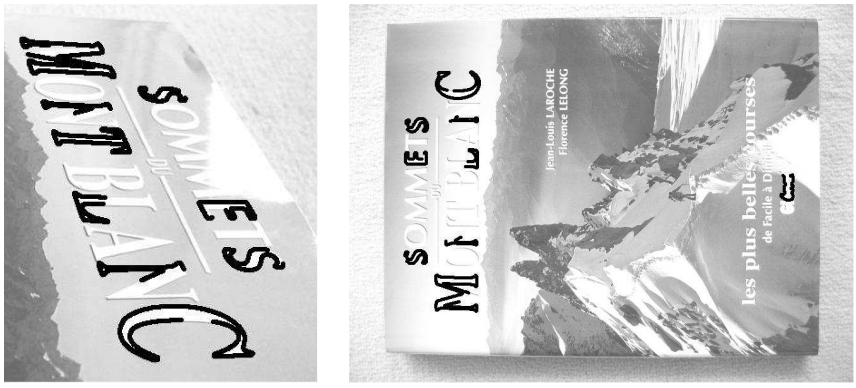
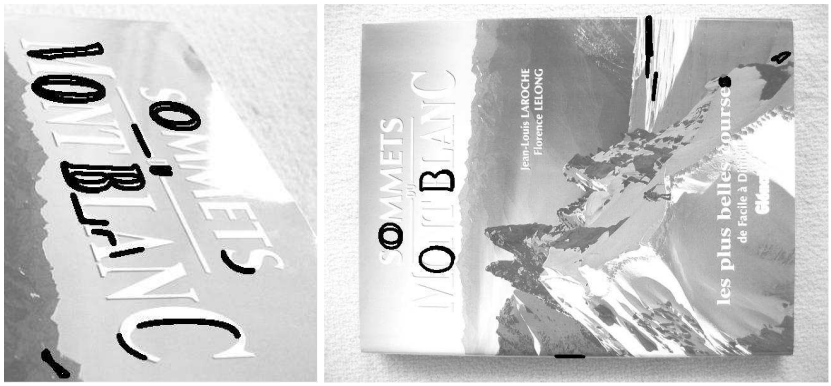


Figure 7.38: Book cover: the 16 semi-local affine invariant matches. The best match has an NFA of 10^{-10}. The scene shape element of the only false match ($NFA = 0.6$) that was detected can be seen in the right part of the scene image

The next stage of the matching procedure consists in finding matches between global shape elements, extracted from those maximal meaningful boundaries that were not described by any semi-local shape element. All not semi-locally encoded maximal meaningful boundaries are shown in Figure 7.39(a). These two sets of curves are used as the input of the global affine invariant recognition method. Figure 7.39(b) shows the detected 1-meaningful matches between global shape elements. Good matches reach NFAs as low as 10^{-10}. Some false matches were detected, but it can only be said they are false because, semantically, they do not correspond

to the same objects. However, these "false matches" correspond to global shapes elements that look actually alike. Such "false" correspondences can often occur: convex or "quasi-convex" shapes are indeed not very discriminatory. Higher level information (such as spatial coherence between matches) is needed in order to assess their semantic validity.

(a)



(b)

Figure 7.39: Book cover. (a) all not locally encoded maximal meaningful boundaries. Too small or too convex level lines are not encoded. (b) the 160 matches between global shape elements, using the global affine invariant recognition method. The search is only performed upon the maximal meaningful boundaries which were not locally encoded. The NFA of some matches reach values as low as 10^{-10}. Since spatial coherence between matches is not taken into account, "false" matches (false from a semantic viewpoint) are unavoidable (these matches correspond to global shapes elements that look actually alike)

Notice that if we combine the matches that were obtained from both the semi-local and affine invariant methods, almost all shapes in common have been detected. Compare now the combination of these matches with the matches detected when using the global method over all the maximal meaningful boundaries (Figure 7.40). It can be observed that using first the semi-local method, and then the global method over the non semi-locally encoded boundaries, produces less false matches than using the global method over the original sets of maximal meaningful boundaries. Indeed, even when not dealing with occlusions, considering semi-local descriptions for "complex" boundaries is more sound than describing them globally, since it allows to increase the discriminatory power.

Figure 7.40: Book cover: the 857 global shape elements detected as 1-meaningful matches, among all maximal meaningful boundaries. The lowest NFA reaches 10^{-14}. The majority of the "false" matches are unavoidable, since globally, the matched shape elements are very alike

Two frames of a sequence

Figure 7.41 shows the semi-locally matched shape elements between two frames of a sequence, using the semi-local similarity invariant method. The non semi-locally encoded maximal meaningful boundaries are displayed in Figure 7.42. The majority of the non semi-locally encoded boundaries are oval shaped, and not discriminatory enough to decide if a match is "semantically correct". Nevertheless, while pairing two of them may not provide much information, looking for spatial coherence between all pairs of matches can lead to high confidence detections.

Figure 7.43 shows some global matches (those for which the NFA is below 10^{-2}). Among the represented shape elements, almost all of them seem to be discriminatory enough, and no "oval" shaped (not discriminatory) boundary is present. This fact is consistent with one of the features of the proposed detection methodology: good matches between discriminatory shape elements show the lowest NFAs.

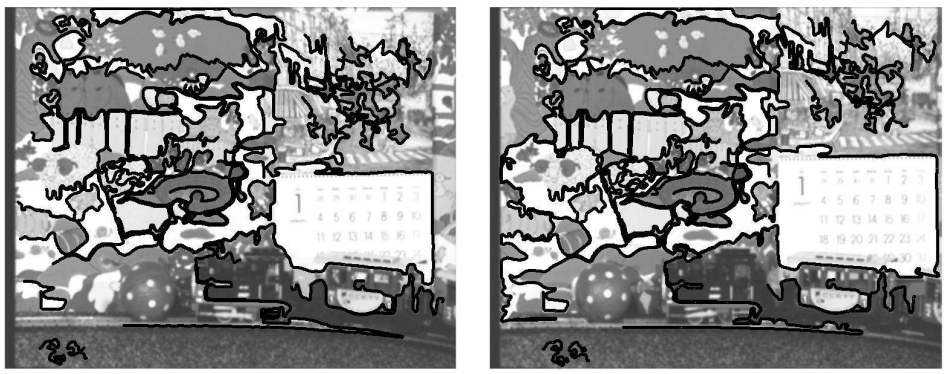


Figure 7.41: Movie frames. The 75 semi-local similarity invariant 1-meaningful matches. The lowest NFA is about $2.0\,10^{-16}$

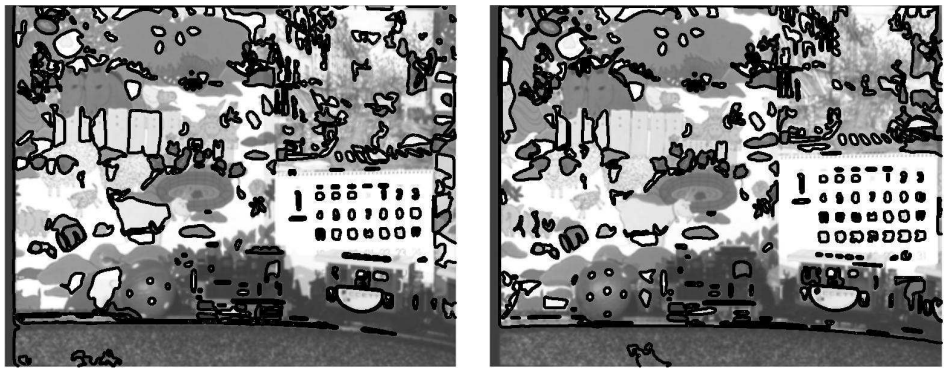


Figure 7.42: Movie frames. Non semi-locally encoded maximal meaningful boundaries. There are 356 lines in the query image (left) and 373 in the scene image (right)

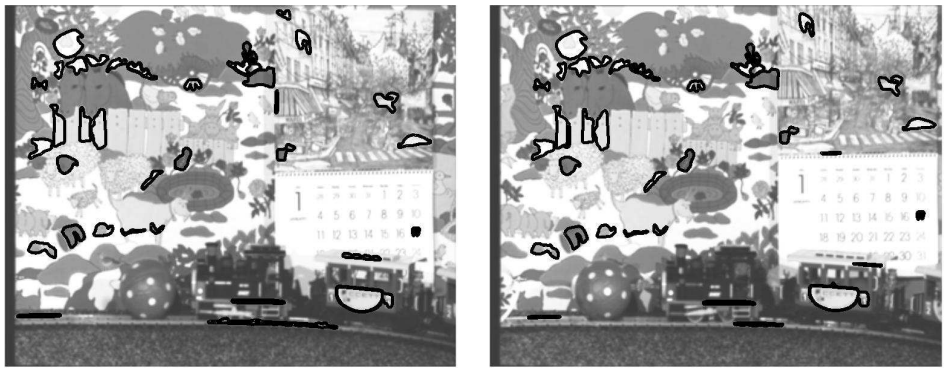


Figure 7.43: Movie frame. The 120 global 10^{-2}-meaningful matches among the non semi-locally encoded level lines. The lowest NFA is about $5.0\,10^{-13}$

Part IV

Grouping shape elements

Chapter 8

Hierarchical clustering and validity assessment

he unsupervised classification of points into groups is commonly referred to as *clustering* or *grouping*. Clustering aims at discovering structure in a point data set, by dividing it into its "natural" groups. There are three classical problems related to the construction of the right clusters. The first one is to evaluate the *validity* of a cluster candidate. In other words, is a group of points really a cluster, i.e. a group with a large enough density? The second problem is that meaningful clusters can contain or be contained in other meaningful clusters. A *rule* is needed to define locally optimal clusters by inclusion. This rule, however, is not enough to interpret correctly the data. The third problem is the definition of a correct merging rule between meaningful clusters, permitting to decide whether they should stay separate or unit. A unified *a contrario* method will be proposed for these problems. In continuation, some complexity issues and heuristics to find sound candidate clusters will be considered.

8.1 Clustering analysis

The previous chapters proved that it is possible to define shape elements in images, with invariance properties that agree with visual perception. This definition is accurate in the sense that two random shape elements look similar with a very small probability in the *ad hoc* background model. This is, however, only a first stage in the shape identification process. The next stage should assert whether several shape elements belong to a same shape, or not. Shape elements must be grouped into what would be more properly called a shape. In this chapter, the grouping problem will be addressed in a very general setting, where the problem is to group data points in a metric space. The next chapter will develop the particular application to shape element grouping. The point data set will then be specific. Each point will be a geometric transform (similarity or affine transform) predicted by a matching pair of shape elements. Each cluster of transforms will correspond to a globally recognized shape.

The classification of general data into groups is usually referred to as clustering. Let $E \subset \mathbb{R}^D$ and consider a data set $\mathcal{D} = \{x_1, \ldots, x_M\}$ of M points in E. (Some of them may be equal.) The clustering problem consists in finding disjoints groups G_1, ..., G_k with $\cup_{i=1}^{k} G_i \subset \mathcal{D}$. The inclusion can *a priori* be strict; the G_i may not form a partition of $\mathcal{D}$. Of course, in order to give a quantitative relevance to each group, E is equipped with a dissimilarity function $d : E \times E \to \mathbb{R}_+$. The groups are then constructed so that each one contains homogeneous data (intra-cluster similarity), and the content of different groups are fairly different (inter-cluster dissimilarity). Many techniques achieving this goal have been proposed. The reader is referred to the Keynotes A.2 for a short review and useful references. The class of method to be used depends on the problem and the type of data to be processed.
Still, there are three main general problems associated with cluster detection, that are also illustrated on Fig. 8.1:

1. Cluster validity: how to assess the relevance of a group of data points? A validity, or meaningfulness measure should be defined.
2. Optimization: how to find relatively exact borders for each group?
3. Merging rule: when two valid clusters are included in another one, is it better to merge them or to keep them separate?

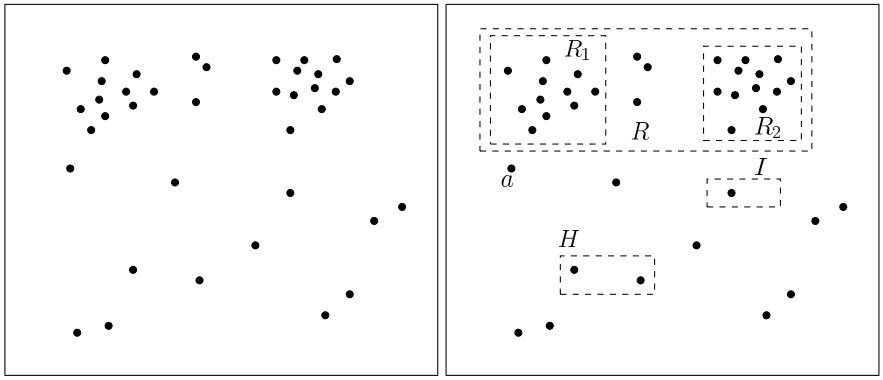


Figure 8.1: This figure illustrates three aspects of the grouping problem. The figure presents a set of data points in the plane and some test regions where an exceptional density may be observed, or not. Intuitively, the regions H and I do not contain clusters. So a first question is to rule out such non meaningful clusters. A second question is the choice of sound candidate regions: for instance, should not R_1 be enlarged to include the point a? As a last question, what is the best description of the observed clusters? The region R is a possible good candidate, but it also contains the points of regions R_1 and R_2 which also are sound candidates. Thus, the question arises of whether R should be chosen as cluster region, rather than the pair (R_1, R_2).

This chapter describes an *a contrario* decision method to answer these three questions.

8.2 *A contrario* cluster validity

8.2.1 The background model

In all what follows, $E \subset \mathbb{R}^D$ is endowed with a probability measure π (which will be also called *background law*.) By definition, $\pi(R)$ is the probability that a random point belongs to R. We do not mention measurability issues here. They are straightforward in this context.
The definition of π is problem specific. In general, it is given *a priori*, or can be empirically estimated over the data, as will be detailed in this chapter and in the following for the shape identification problem.

DEFINITION 8.1 *A* background process *is a finite point process $(X_i)_{i=1,\,...\,M}$ in E made of M mutually independent variables, identically distributed with law π.*

A standard way to construct such a point process from (E, π) is to consider the product probability space $(E^M, \Pr = \pi^M)$ and the random variables X_i defined by $X_i(x) = x_i$ for any $x = (x_i)_{i=1,\,...,\,M} \in E^M$.

Let us now consider an observed data set of M points $\{x_1, \ldots, x_M\}$ in E. Exactly as in Chap. 6, a subset of the data set will form a meaningful group if it could not "occur by chance". In other words, it could not be explained by the background model. Therefore, the cornerstone of the *a contrario* method is to contradict the following assumption.

(A) *The observed M-tuple $(x_i)_{i\in\{1...M\}}$ is a realization of the background process.*

Let us give an example to illustrate this idea. Figure 8.2 represents two 2D projections of a 4-dimensional set of points. These points correspond to similarities applying a shape element in an image to the matched shape

element in another image by the method described Chapter 9. The "high density" of a region of the space reveals that the points therein correspond to the same shape. The probability that such a concentrated cluster be a realization of the background process is intuitively very low.

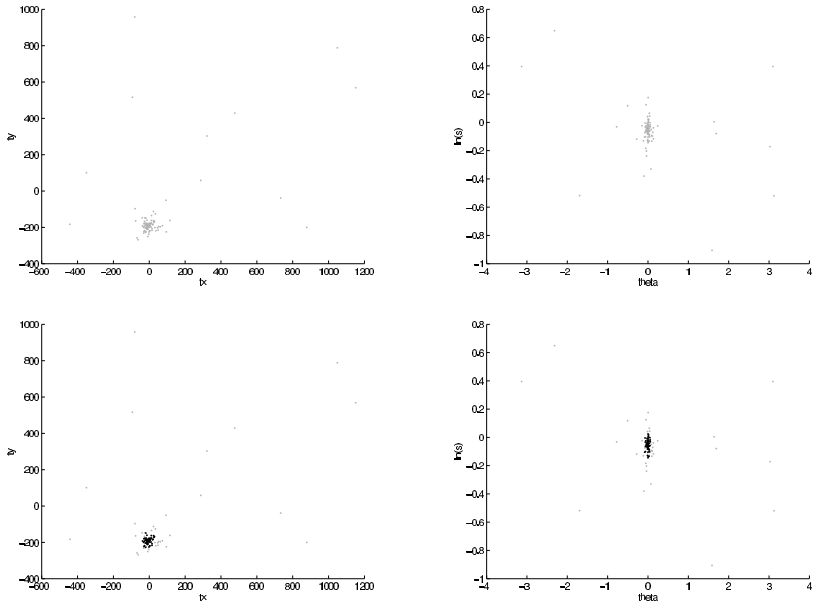


Figure 8.2: Two projections of a 4-dimensional data point set corresponding to a problem studied in Chapter 9, Fig. 10.16. Each dot represents a similarity associated with a meaningful match between shape elements. A group of dots corresponds to a coherent set of similarities indicating that the matched shapes belong to the same global shape. Thus, optimally assembling shapes reduces to the search of optimal point clusters in dimension 4. The question of finding the right groups is crucial. Errors can lead to add spurious elements, or to remove correct elements from a shape. The two top plots are the raw data to be clustered. The two plots below depicts (in black dots) the only group detected by the method exposed in this chapter.

It is assumed that an agglomeration algorithm is given. This is defined as a function

$$\begin{array}{rccl} \mathcal{A}: & E^M & \rightarrow & (\mathcal{P}(E))^P \\ & (x_1,\ldots,x_M) & \rightarrow & \mathcal{A}(x_1,\ldots,x_M) = (G_1,\ldots,G_P) \end{array} \tag{8.1}$$

which to any M-tuple of data points associates a P-tuple of sets, G_1, ..., G_P, such that each G_k is a part of $\{x_1,\ldots,x_M\}$. The algorithm $\mathcal{A}$ is designed from any clustering algorithm and proposes a set of groups candidates from a set of data points. The number of candidates P only depends on the number of data points M and not on the particular values of x_1, ..., x_M. A particular choice for $\mathcal{A}$ will be given in Sect. 8.4, but all the theory below does not depend on this choice. Some of the candidate groups may be actually empty, meaning that P is an upper bound of the number of possible groups.

8.2.2 Meaningful groups

Consider a small region $R \subset E$ containing the origin, typically a hyperrectangle centered at the origin. Assume that k points among $(x_1, \ldots, x_M)$ belong to a region of the type $x_j + R$, for some j, $1 \leqslant j \leqslant M$. If k is large enough, and R small enough, one will observe a cluster of points in R which can hardly have been generated by the background model. This group of points will then be detected *a contrario* in $x_j + R$. Clusters can be grouped around any of the x_j and can have any shape. A generic shape for the tested regions must, however, be

fixed *a priori*. The region R will have to belong to a finite family $\mathcal{R}$ of regions, which will be detailed further. For the time being, let us simply assume that $\mathcal{R}$ has finite cardinality $\#\mathcal{R}$ and that $0 \in R$ for all $R \in \mathcal{R}$.

In the following, for $k \leqslant M \in \mathbb{N}$ and $0 \leqslant p \leqslant 1$, let us denote the tail of the binomial law by

$$\mathcal{B}(M,k,p) = \sum_{j \geqslant k} \binom{M}{j} p^j (1-p)^{M-j}.$$

Given a background process $X_1, \ldots, X_M$ and a region R of E with probability $\pi(R)$, $\mathcal{B}(M,k,\pi(R))$ can be interpreted as the probability that *at least k out of the M points of the process belong to R*. A thorough study of the binomial tail and its use in the detection of geometric structures is presented in [47].

DEFINITION 8.2 *Let $G \subset \{x_1, \ldots x_M\}$ a subset of k points out of the M data points. We call number of false alarms of G,*

$$NFA_g(G) \equiv \#\mathcal{R} \cdot M \cdot P \cdot \min_{\substack{x_j \in G, R \in \mathcal{R} \\ G \subset x_j + R}} \mathcal{B}(M-1, k-1, \pi(x_j + R)). \tag{8.2}$$

We say that G is an ε-meaningful group if $NFA_g(G) < \varepsilon$.

As a sanity check of the above definition, the aim is to prove that the expected number of ε-meaningful regions is less than ε, when the data set x_1, ..., x_M is a realization of the background process and the group candidates result from the agglomeration algorithm $\mathcal{A}$.

A careful notation is needed. Let us fix $1 \leqslant j \leqslant M$ and $R \in \mathcal{R}$. We note:

- $X = (X_1, \ldots, X_M)$, the background process,
- $x = (x_1, \ldots x_m)$ a set of M points in E,
- $X^j = (X_1, \ldots, X_M)$ with X_j omitted in the list,
- $x^j = (x_1, \ldots, x_M)$ with x_j omitted in the list,
- $d\pi^j(x^j) = d\pi(x_1) \ldots d\pi(x_M)$ with $d\pi(x_j)$ omitted in the product,
- $\Pr^j$ the marginal of $\Pr$ with respect to X^j,
- $K(X^j, X_j, R)$, number of points in the list X^j belonging to $X_j + R$.

LEMMA 8.1 *Let us fix $x_j \in E$. Consider a random process X_1, ..., X_M. Then*

$$\Pr{}^j \left(\mathcal{B}(M-1, K(X^j, x_j, R), \pi(x_j + R)) < \frac{\varepsilon}{\#\mathcal{R} \cdot M} \right) \leqslant \frac{\varepsilon}{\#\mathcal{R} \cdot M}.$$

Proof: The repartition function of the random variable $K(X^j, x_j, R)$ is $k \to \mathcal{B}(M-1, k, \pi(x_j + R))$. The results follows from Lem. 3.2, p. 17. □

PROPOSITION 8.1 *Let X_1, ... X_M be a background process. Consider the P random groups $\mathcal{A}(X_1, \ldots, X_M) = (\Gamma_1, \ldots, \Gamma_P)$. Then the expected number of the ε-meaningful groups among Γ_1, ..., Γ_P is less than ε.*

Proof: Let us note

- For $1 \leqslant i \leqslant P$, the Bernoulli variable

$$Y_i = \begin{cases} 1 \text{ if } \Gamma_i \text{ is } \varepsilon\text{-meaningful,} \\ 0 \text{ otherwise.} \end{cases}$$

- $S = \sum_i Y_i$ the number of ε-meaningful groups.

Let us also denote by K_i the (random) cardinality of Γ_i and $\epsilon = \frac{\varepsilon}{MP\#\mathcal{R}}$.

$$\begin{aligned}
\Pr(Y_i = 1) &= \Pr\left(\min_{\substack{X_j \in \Gamma_i, R \in \mathcal{R} \\ \Gamma_i \subset X_j + R}} \mathcal{B}(M-1, K_i - 1, \pi(X_j + R)) < \epsilon \right) && (8.3)\\
&= \Pr(\exists j, R \text{ s.t. } X_j \in \Gamma_i, \Gamma_i \subset X_j + R, \\
&\qquad\qquad \mathcal{B}(M-1, K_i - 1, \pi(X_j + R)) < \epsilon) && (8.4)\\
&\leqslant \Pr(\exists j, R \text{ s.t. } \mathcal{B}(M-1, K(X^j, X_j, R), \pi(X_j + R)) < \epsilon) && (8.5)\\
&\leqslant \sum_{\substack{1 \leqslant j \leqslant M \\ R \in \mathcal{R}}} \Pr(\mathcal{B}(M-1, K(X^j, X_j, R), \pi(X_j + R)) < \epsilon). && (8.6)
\end{aligned}$$

The first inequality results from $\Gamma_i \subset X_j + R \Rightarrow K_i - 1 \leqslant K(X^j, X_j, R)$ and the monotonicity of the map $k \mapsto \mathcal{B}(M-1, k, p)$. Now, Lem. 8.1 cannot be directly applied. Indeed, the considered region is centered at a random point X_j and thus has a random probability. However, by Fubini Theorem

$$\begin{aligned}
&\Pr(\mathcal{B}(M-1, K(X^j, X_j, R), \pi(X_j + R)) < \epsilon) \\
&\qquad = \int d\pi(x_j) \Pr{}^j(\mathcal{B}(M-1, K(X^j, x_j, R), \pi(x_j + R)) < \epsilon), \\
&\qquad \leqslant \int d\pi(x_j)\epsilon \quad \text{by Lem. 8.1}, \\
&\qquad = \epsilon.
\end{aligned}$$

Thus

$$P(Y_i = 1) \leqslant M\#\mathcal{R}\epsilon = \frac{\varepsilon}{P}.$$

Finally,

$$\mathbb{E}(S) = \sum_{i=1}^{P} \mathbb{E}(Y_i) \leqslant \sum_{i=1}^{P} \frac{\varepsilon}{P} = \varepsilon. \qquad \square$$

Remark: As in chapters 3 and 6, the key point is that the *expectation* of the number S of meaningful regions is easily controlled. The probability law of S would instead be extremely difficult to compute because of the interactions between regions.

Let us summarize: the number of false alarms is a measure of how likely it is that a group G containing at least k of data points, was generated "by chance", as a realization of the background process. The lower $NFA_g(G)$, the less likely the observed cluster in the background process. By Prop. 8.1, the only parameter controlling the detection is ε. This provides a handy way to control false detections. If, on the average, one is ready to tolerate one "non relevant region" among all regions, then ε can be simply set to 1.

The following proposition shows that the influence of the parameter $\#\mathcal{R}$ and of the decision parameter ε on the detection results are very weak.

PROPOSITION 8.2 ([47]) *Let* $0 < p < 1$ *and*

$$k^* = \min\{k : MP\#\mathcal{R} \cdot \mathcal{B}(M-1, k, p) \leqslant \varepsilon\}.$$

Then

$$\alpha\sqrt{2p(1-p)} \leqslant k^* - p(M-1) \leqslant \frac{\alpha}{\sqrt{2}}, \tag{8.7}$$

where $\alpha = \sqrt{(M-1)\ln(MP\#\mathcal{R}/\varepsilon)}$.

Notice that k^* is the minimal number of points in a ε-meaningful group, and thus depends on the size of the regions containing the group. By the preceding result, this decision threshold only has a logarithmic dependance upon P, $\#\mathcal{R}$ and ε.

Figure 8.3 shows an example of clustering. The data consists of 950 points uniformly distributed in the unit square, and 50 points manually added around the positions $(0.4, 0.4)$ and $(0.7, 0.7)$. The figure shows the result of a numerical method involving the above NFA. Both visible clusters are found with NFA respectively equal to 10^{-8} and 10^{-7}. Such low numbers can barely be the result of chance. How to obtain *exactly* these two clusters and no other larger or smaller ones which would also be meaningful? This will be the object of the next two sections.

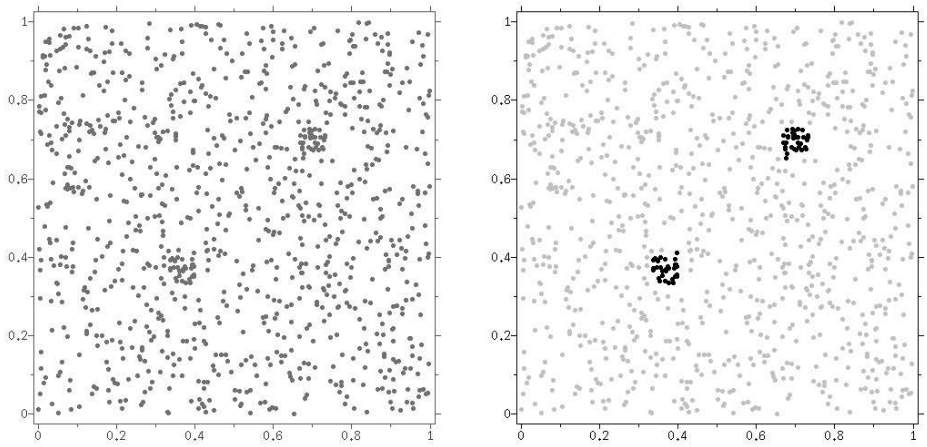


Figure 8.3: Clustering of twice 25 points around $(0.4, 0.4)$ and $(0.7, 0.7)$ surrounded by 950 i.i.d. points, uniformly distributed in the unit square. The regions of $\mathcal{R}$ are rectangles as described in Sect. 8.4.1. In this example $\#\mathcal{R} = 2500$ (50 different sizes in each direction). The distance between data points is the usual Euclidean distance. Exactly two maximal meaningful clusters are detected. The NFA of the lower left one is 10^{-8} while the upper-right one has a NFA equal to 10^{-7}.

8.3 Optimal merging criteria

8.3.1 Local merging criterion

While each meaningful group is relevant by itself, the whole set of meaningful regions exhibits in general a high redundancy. Indeed, a very meaningful group G usually remains meaningful when it is slightly enlarged or shrunk into a group G'.

If, e.g. $G \subset G'$, this question is easily answered by a comparing $NFA_g(G)$ and $NFA_g(G')$. The group with the smallest number of false alarms must of course be preferred. Another more subtle question arises when three or more groups interact. Let G^1 and G^2 be two tested disjoint groups and G another tested group containing $G^1 \cup G^2$. We then face two conflicting interpretations of the data: two clusters or just one? The merged group G is not necessarily a better data representation than the two separate clusters G^1 and G^2. A first possibility is that G is less meaningful than each one of the merging groups. In such a case, G^1 and G^2 should be kept, rather than G. The situation is less obvious when G is more meaningful than both G^1 and G^2. In that case, the merging decision may still be opportune. So a quantitative merging criterion is required. We shall first define a *number of false alarms for a pair of groups*. This new value will be compared to the NFA of the merged group. Let us introduce the trinomial coefficient

$$\binom{M}{i,j} = \binom{M}{i}\binom{M-i}{j}.$$

We note

$$\mathcal{M}(M, k_1, k_2, \pi_1, \pi_2) = \sum_{i=k_1}^{M} \sum_{j=k_2}^{M-i} \binom{M}{i,j} \pi_1^i \pi_2^j (1-\pi_1-\pi_2)^{M-i-j}. \tag{8.8}$$

This number can be interpreted as follows. Let R_1 and R_2 be two disjoint regions of E and $\pi_1 = \pi(R_1)$, $\pi_2 = \pi(R_2)$ their probabilities. Then $\mathcal{M}(M, k_1, k_2, \pi_1, \pi_2)$ is the probability that at least k_1 among the M, and then at least k_2 points among the remaining ones, belong to R_1 and R_2 respectively. Thus, this probability measures how exceptional a pair of concentrated clusters can be in the background model.

As in the case of single regions, it is assumed that a set of P *pairs* of group candidates are obtained by an operator $\mathcal{A}_2$. That is to say

$$\begin{aligned} \mathcal{A}_2 : \quad E^M &\rightarrow (\mathcal{P}(E) \times \mathcal{P}(E))^P \\ (x_1, \ldots, x_M) &\rightarrow \mathcal{A}_2(x_1, \ldots, x_M) = ((G_1^1, G_1^2), \ldots, (G_P^1, G_P^2)), \end{aligned} \tag{8.9}$$

where it is assumed that $G_i^k \subset \{x_1, \ldots x_M\}$, for $k = 1, 2$ and $1 \leqslant i \leqslant P$.

DEFINITION 8.3 *Consider two group candidates (G^1, G^2) of data points. Let $(z_1, z_2) \in G^1 \times G^2$ be two data points, and R_1 and R_2 in $\mathcal{R}$. Let us denote by*

- *k_1 (resp. k_2) the cardinality of $G^1 \backslash (z_2 + R_2)$ (resp. $G^2 \backslash (z_1 + R_1)$), i.e. the number of points of G^1 (resp. G^2) that are not in $z_2 + R_2$ (resp. $z_1 + R_1$).*
- *$\pi_1 = \pi((z_1 + R_1) \backslash (z_2 + R_2))$ and $\pi_2 = \pi((z_2 + R_2) \backslash (z_1 + R_1))$.*

Let us define the number of false alarms of the pair (G^1, G^2) by

$$NFA_{gg}(G^1, G^2) = M^3 \cdot P \cdot (\#\mathcal{R})^2 \min_{\substack{(z_1, z_2) \in G^1 \times G^2, \\ R_1, R_2 \in \mathcal{R} \\ G^1 \subset z_1 + R_1, \\ G^2 \subset z_2 + R_2}} \mathcal{M}(M-2, k_1-1, k_2-1, \pi_1, \pi_2). \tag{8.10}$$

We say that a pair of groups (G^1, G^2) is ε-meaningful if $NFA_{gg}(G^1, G^2) < \varepsilon$.

Let us sum up how to compute this quantity: choose a region centered at one point of G^1 (resp. G^2) and containing G^1 (resp. G^2). Those two regions may intersect, so remove their intersection and the points it may contain. Then, k_1 and k_2 points are left in each group, and the trinomial tail can be computed. As above, this quantity measures how (un)likely it is that G^1 and G^2 contain *respectively* at least k_1 and k_2 points. Removing the intersection is a mere technicality so that the probability of this event is the tail of the trinomial law.

As usual, the aim is to prove that the expected number of ε-meaningful pairs of regions is less than ε. As in the study of ε-meaningful groups, some care must be taken of notations and abbreviations. Let $1 \leqslant i \neq j \leqslant M$. Now, two tested regions $x_i + R_i$ and $x_j + R_j$ may intersect and we have to deal with this possibility. (The indices i and j in R_i and R_j only aim at reminding that these are regions centered at x_i and x_j, although this notation is a bit incorrect.) We note

- $X = (X_1, \ldots, X_M)$, the background process,
- $x = (x_1, \ldots x_M)$ a set of M dots in E,
- $X^{ij} = (X_1, \ldots, X_M)$ with X_i, X_j omitted in the list,
- $x^{ij} = (x_1, \ldots, x_M)$ with x_i, x_j omitted in the list,
- $X_{ij} = (X_1, \ldots, X_M)$ with X_i and X_j replaced by x_i and x_j,
- $d\pi^{ij}(x^{ij}) = d\pi(x_1) \ldots d\pi(x_M)$ with $d\pi(x_i)$ and $d\pi(x_j)$ omitted in the product,

- $\Pr^{ij}$ the marginal of Pr with respect to x^{ij},
- $K(X,i,j,R_i,R_j)$ = the number of points among X^{ij} that are in X_i+R_i but not in X_j+R_j, i.e. belonging to $(X_i+R_i)\setminus(X_j+R_j)$,
- $K_i = K(X,i,j,R_i,R_j)$, $K_j = K(X,j,i,R_j,R_i)$,
- $\tilde{K}_i = K(X_{ij},i,j,R_i,R_j)$, $\tilde{K}_j = K(X_{ij},j,i,R_j,R_i)$,
- $k_i = K(x_i,i,j,R_i,R_j)$, $k_j = K(x_j,j,i,R_j,R_i)$,
- $\pi_i = \pi((x_i+R_i)\setminus(x_j+R_j))$, $\pi_j = \pi((x_j+R_j)\setminus(x_i+R_i))$,
- $\Pi_i = \pi((X_i+R_i)\setminus(X_j+R_j))$, $\Pi_j = \pi((X_j+R_j)\setminus(X_i+R_i))$,
- $\epsilon = \frac{2\varepsilon}{M^3P(\#\mathcal{R})^2}$.

LEMMA 8.2 *For every $x_i, x_j \in E$,*

$$\Pr^{ij}\left[\mathcal{M}(M-2,\tilde{K}_i,\tilde{K}_j,\pi_i,\pi_j) < \epsilon\right] < (M-1)\epsilon.$$

Proof: The proof extends the arguments used for Lem. 3.2, p. 17 to the case of two variables. Remark that this proof is true for discrete variables, since it used that fact that $\tilde{K}_j$ and $\tilde{K}_i$ can only take $M-1$ different values. Indeed,

$$\begin{aligned}
&\Pr^{ij}\left[\mathcal{M}(M-2,\tilde{K}_i,\tilde{K}_j,\pi_i,\pi_j) < \epsilon\right] \\
&\quad= \sum_{(k_i,k_j)|\mathcal{M}(M-2,k_i,k_j,\pi_i,\pi_j)<\epsilon} \Pr^{ij}(\tilde{K}_i = k_i, \tilde{K}_j = k_j) \\
&\quad= \sum_{(k_i,k_j)|\mathcal{M}(M-2,k_i,k_j,\pi_i,\pi_j)<\epsilon} \binom{M-2}{k_i,k_j} \pi_i^{k_i}\pi_j^{k_j}(1-\pi_i-\pi_j)^{M-2-k_i-k_j}.
\end{aligned}$$

Let

$$k_i(\epsilon,k_j) = \inf\{0 \leqslant k \leqslant M-2 | \mathcal{M}(M-2,k,k_j,\pi_i,\pi_j) < \epsilon\},$$

with the useful conventions $\mathcal{M}(M-2,k,k_j,\pi_i,\pi_j) = 0$ and $\binom{M-2}{k,k_j} = 0$ if $k \geqslant M-1-k_j$. The map $k \to \mathcal{M}(M-2,k,k_j,\pi_i,\pi_j)$ being monotone,

$$\mathcal{M}(M-2,k,k_j,\pi_i,\pi_j) < \epsilon \Leftrightarrow k \geqslant k_i(\epsilon,k_j). \tag{8.11}$$

Summarizing and using the definition of $k_i(\epsilon,k_j)$,

$$\begin{aligned}
&\Pr^{ij}\left[\mathcal{M}(M-2,\tilde{K}_i,\tilde{K}_j,\pi_i,\pi_j) < \epsilon\right] \\
&\quad= \sum_{k_j=0}^{M-2} \sum_{k=k_i(\epsilon,k_j)}^{M-2} \binom{M-2}{k,k_j} \pi_i^{k}\pi_j^{k_j}(1-\pi_i-\pi_j)^{M-2-k-k_j} \\
&\quad\leqslant \sum_{k_j=0}^{M-2} \sum_{k=k_i(\epsilon,k_j)}^{M-2} \sum_{l=k_j}^{M-2-k} \binom{M-2}{k,l} \pi_i^{k}\pi_j^{l}(1-\pi_i-\pi_j)^{M-2-k-l} \\
&\quad= \sum_{k_j=0}^{M-2} \mathcal{M}(M-2,k_i(\epsilon,k_j),k_j,\pi_i,\pi_j) < (M-1)\epsilon. \qquad \square
\end{aligned}$$

PROPOSITION 8.3 *Let us consider a background process X_1, ..., X_M and the P random pairs $\mathcal{A}_2(X_1, \ldots, X_M) = ((\Gamma_1^1, \Gamma_1^2), \ldots, (\Gamma_P^1, \Gamma_P^2))$. Then, the expected number of ε-meaningful pairs of regions among them is less than ε.*

Proof: Let us note for $k = 1, \ldots, P$

- The Bernoulli variable
$$Y_k = \begin{cases} 1 \text{ if } (\Gamma_k^1, \Gamma_k^2) \text{ is } \varepsilon\text{-meaningful,} \\ 0 \text{ otherwise.} \end{cases}$$
- $S = \sum_{k=1}^{P} Y_k$ the number of ε-meaningful pairs of regions.

Let us fix k. Let X_i and X_j two points in the process, belonging to Γ_k^1 and Γ_k^2. Let R_i and R_j be two regions in $\mathcal{R}$, such that $\Gamma_k^1 \subset X_i + R_i$ and $\Gamma_k^2 \subset X_j + R_j$. Let also $\hat{K}_i$ the number of points of Γ_k^1 that are not in $X_j + R_j$ and $\hat{K}_j$ the number of points of Γ_k^2 that are not in $X_i + R_i$. Remark that with the notations above, $\hat{K}_i - 1 \leqslant K_i$ and $\hat{K}_j - 1 \leqslant K_j$. Then,

$$\begin{aligned} \Pr(Y_k = 1) &= \Pr(\exists i, j, R_i, R_j \quad \text{s.t.} \quad X_i \in \Gamma_k^1,\, X_j \in \Gamma_k^2, \\ &\qquad \Gamma_k^1 \subset X_i + R_i, \Gamma_k^2 \subset X_j + R_j, \\ &\qquad \mathcal{M}(M-2, \hat{K}_i - 1, \hat{K}_j - 1, \Pi_i, \Pi_j) < \epsilon). \\ &\leqslant \Pr(\exists i, j, R_i, R_j \quad \text{s.t. } \mathcal{M}(M-2, K_i, K_j, \Pi_i, \Pi_j) < \epsilon) \\ &\leqslant \sum_{i,j=1}^{M} \sum_{R_i, R_j} \Pr(\mathcal{M}(M-2, K_i, K_j, \Pi_i, \Pi_j) < \epsilon) \end{aligned}$$

The first inequality results from $\hat{K}_i - 1 \leqslant K_i$ and $\hat{K}_j - 1 \leqslant K_j$ and the monotonicity of the map $(k, l) \mapsto \mathcal{M}(M-2, k, l, p, q)$ with respect to each of its variables. By Fubini theorem,

$$\begin{aligned} &\Pr(\mathcal{M}(M-2, K_i, K_j, \Pi_i, \Pi_j) < \epsilon) \\ &\quad = \int_{E^2} d\pi(x_i) d\pi(x_j) \int_{E^{M-2}} \mathbb{1}_{\{\mathcal{M}(M-2, k_i, k_j, \pi_i, \pi_j) < \epsilon\}} d\pi^{ij}(x^{ij}) \\ &\quad = \int_{E^2} d\pi(x_i) d\pi(x_j) \Pr{}^{ij}(\mathcal{M}(M-2, \tilde{K}_i, \tilde{K}_j, \pi_i, \pi_j) < \epsilon) \\ &\quad < (M-1)\epsilon, \end{aligned}$$

where Lem. 8.2 has been used in the last inequality. Finally,

$$\begin{aligned} \mathbb{E}(S) &= \sum_{k=1}^{P} \mathbb{E}(Y_k) \\ &< \sum_{k=1}^{P} (M-1) M^2 (\#\mathcal{R})^2 \epsilon \\ &\leqslant \varepsilon. \qquad \square \end{aligned}$$

DEFINITION 8.4 (MERGING CONDITION) *Let G^1 and G^2 be two groups and G containing $G^1 \cup G^2$. We say that G is indivisible relatively to G^1 and G^2 if*

$$NFA_g(G) \leqslant NFA_{gg}(G^1, G^2). \tag{8.12}$$

Equation (8.12) represents a crucial test for the coherence of a cluster. If it is not fulfilled, G will not be considered a valid region, as it can be divided into a more meaningful pair of cluster regions. The next lemma will prove useful in speeding up the merging decision.

LEMMA 8.3 *For every k_1 and k_2 in $\{0,\dots,M\}$, such that $k_1+k_2 \leqslant M$ and for every π_1 and π_2 in $[0,1]$ such that $\pi_1+\pi_2 \leqslant 1$,*

$$\mathcal{M}(M,k_1,k_2,\pi_1,\pi_2) \leqslant \mathcal{B}(M,k_1,\pi_1)\cdot\mathcal{B}(M,k_2,\pi_2). \tag{8.13}$$

A proof of the lemma is given in Annex.A.4. We are actually interested in its consequence to follow.

PROPOSITION 8.4 *Let G be indivisible with respect to G^1 and G^2. Let also assume that the regions related to G^1 and G^2 are disjoint. Then*

$$NFA_g(G) < \frac{M}{P}\cdot NFA_g(G^1)\cdot NFA_g(G^2)$$

*Proof:*Let us denote by π_1 and π_2 the probability of the regions attaining the NFA_g of G^1 and G^2. These regions are assumed to be disjoint. By an obvious argument of monotonicity, those regions also attain the minimum of the trinomial law. From (8.10) and Def. 8.4,

$$NFA_{gg}(G^1,G^2) = M^3P(\#\mathcal{R})^2\mathcal{M}(M-2,k_1-1,k_2-1,\pi_1,\pi_2)$$

and

$$NFA_g(G_i) = MP(\#\mathcal{R})\mathcal{B}(M-1,k_i-1,\pi_i),\ i=1,2.$$

By Lem. 8.3, it follows that

$$NFA_{gg}(G^1,G^2) \leqslant M^3P\cdot(\#\mathcal{R})^2\cdot\mathcal{B}(M-2,k_1-1,\pi_1)\mathcal{B}(M-2,k_2-1,\pi_2).$$

Since $\mathcal{B}(M-2,k-1,p) \leqslant \mathcal{B}(M-1,k-1,p)$ for all M, k and p, the result follows. □

Proposition 8.4 is useful from the computational viewpoint, since in many cases one can avoid computing the tail of the trinomial distribution by "filtering" those clusters that do not pass the necessary condition.

8.4 Computational issues

8.4.1 The choice of test regions

What is the right set of test regions $\mathcal{R}$? This question is obviously application driven. To fix ideas, let us just indicate a possible choice. For some reasonably fixed $a > 0$, $r > 1$ and $n \in \mathbb{N}$, let us consider all hyperrectangles whose edge lengths belong to the set $\{a,\ ar,\ ar^2,\ \dots,\ ar^n\}$. This allows one to consider a tractable number of test regions with very different sizes and shapes. The choice of the hyperrectangles is particularly opportune when the probability distribution π, defined on a hyperrectangle E of $\mathbb{R}^D$, is a tensor product of one-dimensional densities $\pi_1,\ \dots,\ \pi_D$. We address the question with more details in the next chapter, where the distribution π cannot be assumed separable.

Definition 8.2 permits to compute the NFA of any group of points. This computation involves a region centered at a data point. Since the number of scales is n in each dimension, there are n^D possible regions of different size. Each region can be centered at M different data point, which makes Mn^D regions. In the next chapter, $D = 4$ or 6. From the numerical feasibility viewpoint, Mn^D becomes too large when n grows. Hence, detection cannot be performed by scanning all the regions centered at a point, counting the number of points it contains and compute the tail of the binomial law. The agglomeration procedure $\mathcal{A}$ involved in the definition of the meaningful groups precisely aims at reducing the number of test groups. In the experiments that follow, this agglomeration algorithm is classical hierarchical clustering algorithm. It provides a binary tree structure from a data set $(x_1,\dots,x_M)$. This tree, sometimes called *dendrogram*, contains exactly $2M-1$ nodes, M of

which are singletons. The pairs of $\mathcal{A}_2$ are simply obtained as the children of a node of this binary tree. There are at most $P = M - 1$ of them.

More generally speaking, hierarchical clustering methods provide a family of nested partitions of the point data set, that can always be embedded in a tree structure (that may not be binary).

Sect. A.2 describes some of the main aggregation techniques to build such trees. Many of them proceed by a recursive binary merging procedure. Thus, they directly yield binary trees. In such methods, the initial set of nodes is the set of data singletons, $\{x_1\}, \ldots, \{x_N\}$. At each stage of the construction, the two closest nodes are united to form their parent node. The inter-cluster distance must be chosen *ad hoc*. In the case of sparse data, one can take the minimal distance $d(x_i, x_j)$ where x_i belongs to the first cluster and x_j to the second one. The nodes of the tree are all merged parts at all levels and the children of a node are the two parts it was merged from. Let us point out that the result much depends on the choice of the dissimilarity function between clusters, for which there is no universal choice.

Such a construction introduces some arbitrariness, but it can become necessary. Indeed, the set of all possible partitions of a data point set is huge. A tree structure permits to reduce the exploration to the search of an optimal subtree of the initial tree structure. This reduction makes sense if the set of nodes of the initial tree structure contains roughly all groups of interest. The choices of the right metric on the data point set and of the right inter-cluster distances are therefore crucial.

Given a dendrogram of the data point set, the following algorithm permits to explore all regions centered at data points and containing a node of the dendrogram.

Grouping algorithm
For each node G (candidate group) with cardinality k in the clustering tree or dendrogram.

1. Set $NFA(G) \leftarrow +\infty$.
2. For $x \in G$,
   (a) Find the smallest region $x + R$ centered at x, and containing the other data points of the node.
   (b) Set $NFA(G) \leftarrow \min(NFA(G), MP \cdot \#\mathcal{R} \cdot \mathcal{B}(M-1, k-1, \pi(x+R)))$.

8.4.2 Indivisibility and maximality

We are now faced with Questions 2 and 3 mentioned at the beginning of the present chapter: we can get many meaningful clusters by the preceding method. Their NFA is known. One can also compute the NFA of a pair of clusters, and compare it roughly to the NFA of their union. The next definition proposes a way to select the right clusters, by using the cluster dendrogram.

DEFINITION 8.5 (MAXIMAL ε-MEANINGFUL GROUP) *A node G is maximal ε-meaningful if and only if*

1. $NFA_g(G) \leqslant \varepsilon$,
2. *G is indivisible with respect to any pair of sibling descendents,*
3. *for all indivisible descendent G', $NFA_g(G') \geqslant NFA_g(G)$,*
4. *for all indivisible ascendent G', either $NFA_g(G') > NFA_g(G)$ or there exists an indivisible descendent G'' of G' such that $NFA_g(G'') < NFA_g(G')$.*

Condition 4 implies that G can be abandoned for a larger group only if this group has not been beaten by one of its descendents. Imposing conditions 3 and 4 ensures that two different maximal meaningful groups are disjoint.

Let us illustrate the critical importance of the merging condition with two simple examples. Figure 8.4 shows a configuration of 100 points, distributed on $[0,1]^2$, and naturally grouped in two clusters G^1 and G^2. In the hierarchical structure, G^1 and G^2 are the children of $G = G^1 \cup G^2$. All three nodes are obviously

meaningful, since their NFA_g is much lower than 1. Their NFA_g also is lower than the NFA_g of the other groups in the dendrogram. Taking a uniform background law, it has been checked that for this particular configuration,

$$NFA_g(G^2) < NFA_g(G) < NFA_g(G^1).$$

It is clear that G^1 represents an informative part of the data that should be kept. This will be the case. Notice that G^2 is more meaningful than G and is contained in G. Thus, G would be eliminated if only the most meaningful groups by inclusion were kept. On the other hand, G is more meaningful than G^1, so that G^1 is not a local maximum of meaningfulness, with respect to inclusion. So, without the notion of indivisibility, trouble would arise: G would eliminate G^1 and G^2 would eliminate G. The result would be the solution indicated in the middle column of Fig. 8.4. Actually G is neither indivisible nor satisfies Condition 3 above, since it is less meaningful than the pair (G^1, G^2) and than G^2. Thus, the result of the grouping procedure yields, in accordance with the rule of Def. 8.5, the pair (G^1, G^2).
In [50], the above mentioned maximality definition was proposed: it consists of taking the lowest NFA in all the branches of the tree. As has just been seen, this definition is not suitable here. By this definition, G^1 would have been considered as the only maximal meaningful cluster of the tree.

Fig. 8.5 illustrates another situation where the indivisibility check yields the intuitively right solution. In this example, the union G of two clusters G^1 and G^2 is more meaningful than each separate cluster. Without the indivisibility requirement, G would be the only maximal meaningful group. This would have been coherent, had G^1 and G^2 been intricate enough. In the presented case, the indivisibility condition yields two clusters G^1 and G^2, since $NFA_{gg}(G^1, G^2) < NFA_g(G)$.

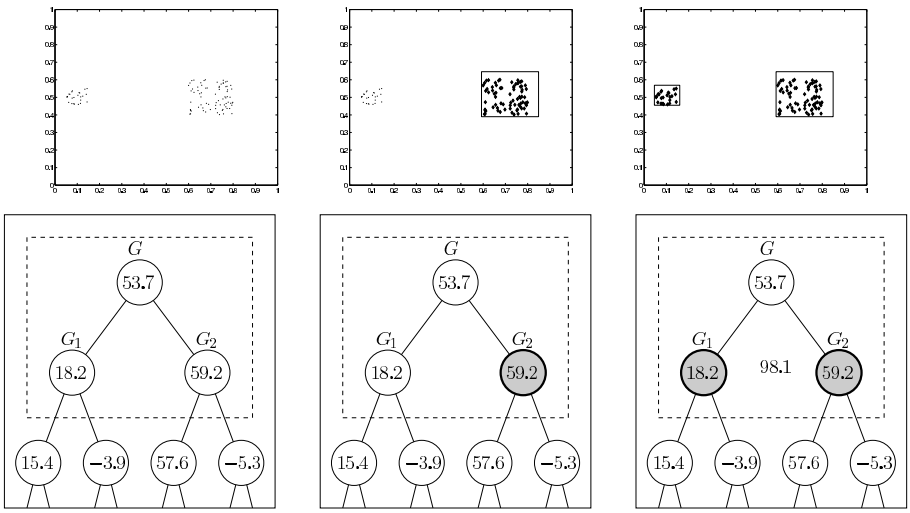


Figure 8.4: Indivisibility prevents collateral elimination. Each subfigure shows a configuration of points, and a piece of the corresponding dendrogram, with the selection of maximal meaningful groups, depicted in grey. The number in each node corresponds to $-\log_{10}(NFA_g)$ of its associated cluster, so that the cluster is meaningful when this number is large. The number placed between two nodes is the NFA_{gg} of the corresponding pair. Left: original configuration. Middle: the node selected by taking only the most meaningful group in each branch. The left-most group G^1 is eliminated. It is, however, very meaningful since $NFA_g(G^1) = 10^{-18}$. Right: by combining indivisibility and maximality criteria, both clusters G^1 and G^2 are selected.

8.5 Experimental validation: object grouping based on elementary features

Grouping phenomena are essential in human perception, since they are responsible for the organization of information. In vision, grouping has been especially studied by Gestalt psychologists like Wertheimer [165].

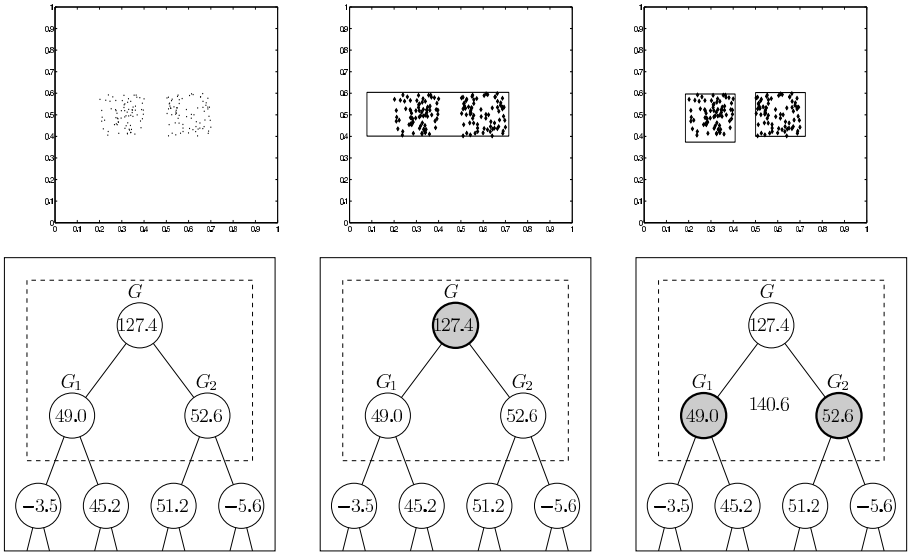


Figure 8.5: Indivisibility prevents faulty union. Each sub-figure shows a configuration of points, and a piece of the corresponding dendrogram, with the selection of maximal meaningful groups, depicted in grey. The number in each node corresponds to the NFA_g of its associated cluster. The number between two nodes is the NFA_{gg} of the corresponding pair. Left: original configuration. Middle: the node selected if one only checks maximality by inclusion and not indivisibility. The largest group G has the lowest NFA_g and would be the only one kept. Note that the optimal region is not symmetric, since it must be centered on a datapoint. Right: selected nodes obtained by combining the indivisibility and maximality criteria. Since $NFA_{gg}(G^1, G^2) = 10^{-140} < 10^{-127} = NFA_g(G)$, the pair (G^1, G^2) is preferred to G.

The aim of these experiments is to extract the groups of objects in an image, that share some elementary geometrical properties. The objects boundaries are extracted as some contrasted level lines in the image, called *meaningful level lines* (see [48] for a full description of this extraction process). Once these objects are detected, say $O_1, ...O_M$, we can compute for each of them a list of D features (grey level, position , orientation, etc...). If k objects among M have one or several features in common, we wonder if it is happening by chance or if it is enough to group them. Each data point is a point in a bounded subset of $\mathbb{R}^D$ and the method described above is applied. (Actually, some coordinates, as angles, belong to the unit circle, since periodicity must be taken into account. This can be done all the same.)

8.5.1 Dots in noise

The first experiment is Fig. 8.3, which contains two groups of 25 points in addition to 950 i.i.d uniformly in the unit square. Two groups and two groups only are detected with very good NFA_g (less than 10^{-7}).

8.5.2 Segments

In the second example, groups are perceived as a result of the collaboration between two different features. Figure 8.6 shows 71 straight segments with different orientations, almost uniformly distributed in position. As expected, no meaningful cluster is detected in the space of position coordinates of the barycenters. In all the experiments, the number of rectangle sizes in each direction is 50. Thus $\#\mathcal{R} = 50^D$.

If orientation is chosen as the only feature ($D = 1$), 6 maximal meaningful groups are detected, corresponding to the most represented orientations. None of these clusters exhibits a very low NFA_g. The central group is not even detected, because the directions of the segments are slightly different. It only means that orientation is not the only perceptual grouping law used in the interpretation of this figure. All the other groups

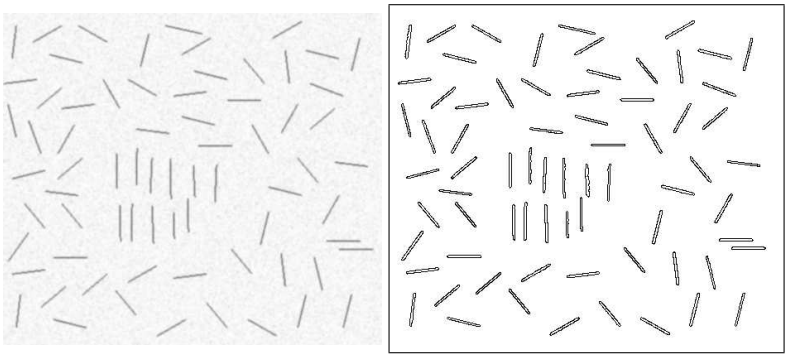


Figure 8.6: An image of a scanned drawing of segments, and its 71 maximal meaningful level lines [48].

are actually not perceived, because they are masked by the clutter made of all the other objects. However, one cannot deny that they have a coherent direction.

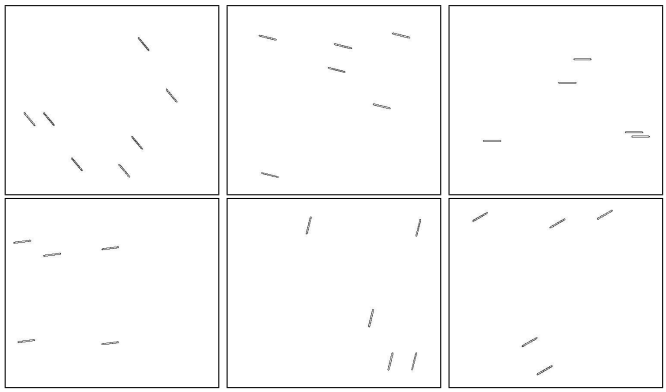


Figure 8.7: Grouping with respect to orientation: there are 6 maximal meaningful groups. NFA_g range is between $10^{-0.4}$ and $10^{-3.8}$. Remark that the central group is missing. Indeed, the direction of the segment is not accurate, and the group is not meaningful with respect to orientation. This experiment shows that orientation alone is not sufficient to detect some groups. Orientation is only one law of perceptual grouping among others.

Now, let us see what happens when considering two features ($D = 2$, $\#\mathcal{R} = 2500$). In the space (x-coordinate, orientation), two maximal meaningful clusters are found (Fig. 8.8). As expected, the most meaningful is the group G of 11 central vertical segments. Its NFA_g is equal to $10^{-1.5}$, which is not that low. The second one is correct, but hardly meaningful $NFA_g = 0.3$. In the space (y-coordinate, orientation), the central group G is splitted into two maximal meaningful clusters. They correspond to the two rows of segments composing G. The role of the merging criterion is decisive here. In the space (y-coordinate, orientation), the combination of the maximality and the merging criterion yields that it is more meaningful to observe at the same time the two rows of segments than the whole G. This is coherent with the visual perception, since we actually see two lines of segments here. On the contrary, in the (x-coordinate, orientation) space, the merging criterion indicates that observing G is more meaningful than observing simultaneously its children in the dendrogram. This decision is still conform with observation: no particular group within G can be distinguished with regards to the x-coordinate. The same group is obtained in the space (x-coordinate, y-coordinate, orientation), with a lower $NFA_g = 10^{-3.4}$.

8.5.3 DNA image

The 80 objects in Fig. 8.9 are more complex, in the sense that more features are needed in order to represent them (diameter, elongation, orientation, *etc.*). It is clear that a projection on a single feature is not really enough to differentiate the objects. Globally, we see three groups of objects: the DNA marks, which share the same

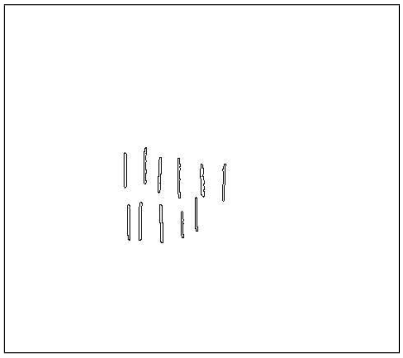


Figure 8.8: Grouping in the space (x-coordinate,orientation). This time, the whole central group is detected and is the only maximal meaninful group. ($NFA_g = 10^{-0.3}$). If grouping is done with respect to full 2D-position and orientation, the central group is still the only detected group with $NFA_g = 10^{-2.3}$.

form, size and orientation; the numbers, all on the same line, almost of the same size; finally the elements of the ruler, also on the same line and of similar diameters. The position appears to be decisive in the perceptive formation of these groups.

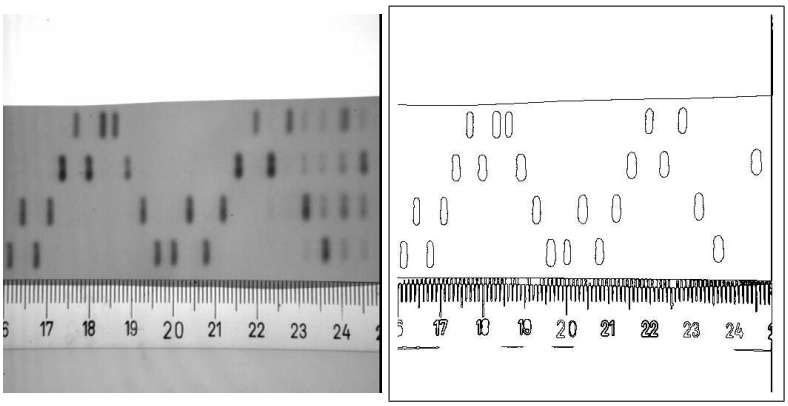


Figure 8.9: An image of DNA and its 80 maximal meaningful level lines [48].

In the space (diameter, y-coordinate), 6 maximal meaningful groups are detected (Fig. 8.10). Four of them correspond to the lines of DNA marks (from left to right and top-down), $-\log_{10}(NFA_g) = 1.2, 6.1, 5.1, 4.3$. The group of numbers contains 23 objects (a group of two digits sometimes contains three objects: the two digits and a level line surrounding both of them) and $-\log_{10}(NFA_g) = 41.7$. The last group, composed of the vertical graduation of the ruler contains 31 objects and is even more meaningful, $-\log_{10}(NFA_g) = 57.3$.

Now, let us give up considering the position information. Do we still see the DNA marks as a group? By taking several other features into account (see Fig. 8.11), the DNA marks form an isolated and very meaningful group: the combination of features (orientation, diameter, elongation, convexity coefficient) reveals the DNA marks as a very good maximal meaningful cluster ($NFA_g = 10^{-10}$). There are two other interesting groups that are not detected, but whose NFA_g is also close to 1: the 1's and the 2's of the ruler. Let us detail how π, the law of the background model was estimated on the data itself: the marginal distribution of each characteristic is approximated by the empirical histogram. Then all the characteristics are assumed to be independent. Let us point out that the obtained distribution is not uniform at all. Why making such a construction, and not taking a uniform law? First, it would assume that the range of the data is known, which is not true. Moreover, each marginal distribution of the different characteristics has no a priori reason to be uniform. Hence, contradicting the background model could actually be due to the lack of independence of the data points, or the discrepancy of their distribution with respect to the uniform distribution. Therefore, it is useful to define a background law which is as realistic as possible for a single observation. Then, why not taking the joint empirical law? Because

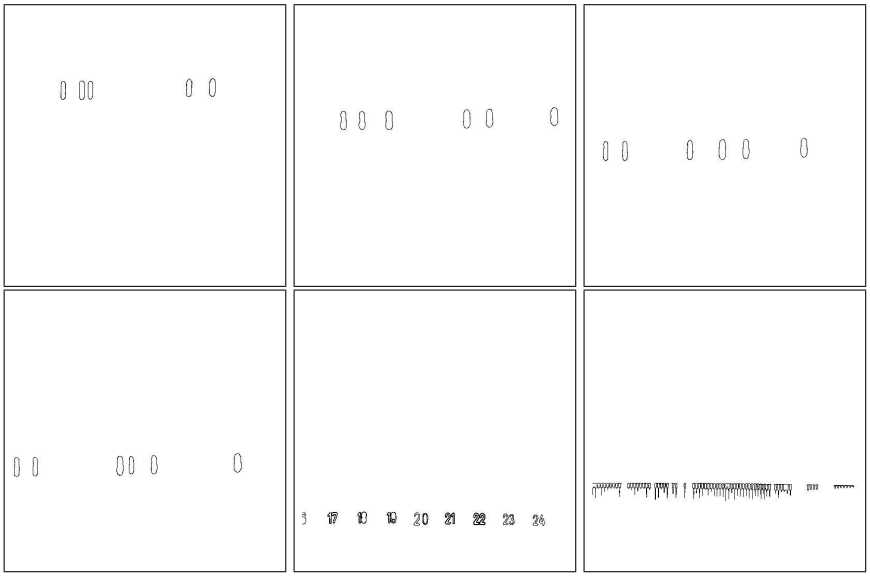


Figure 8.10: Grouping with respect to diameter and y coordinate. Six groups are detected, 4 of which are rows of DNA marks. The last two ones correspond to the ruler. $-\log_{10}(NFA_g)$ range from 1.2 to 6.1 for the DNA. The last two groups are larger and are obviously more meaningful: $-\log_{10}(NFA_g) = 41.7$ and 57.3.

of dimensionality, it is not possible to estimate this distribution, unless thousands or millions of data points are given, which is not the case. Estimating one dimensional marginal laws is less sensitive to this problem of dimensionality. Moreover, the aim is to detect clusters located in small regions. If the true distribution is smooth enough, it is still possible to locally approximate this distribution as the product of marginal laws.

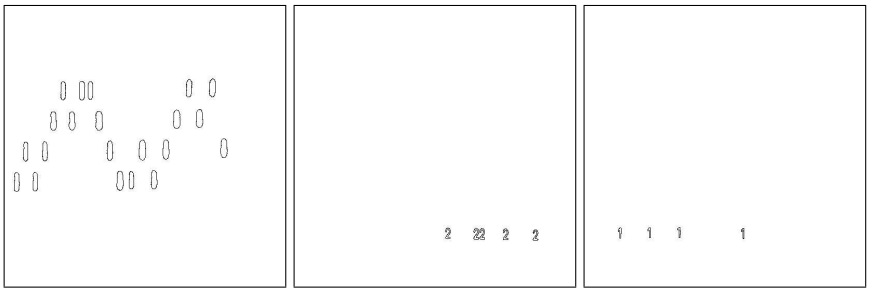


Figure 8.11: Grouping with respect to orientation, elongation, diameter, and a convexity coefficient. The DNA marks are the most meaningful group $NFA_g = 10^{-10}$. It is worth noting that the 1 and 2's, though not meaningful groups, have NFA_g only slightly larger than 1 (1.6 and 4.5).

8.6 Bibliographical notes

The problem of finding groups in a large data set is an active research field. It is involved in data-mining, pattern recognition and pattern classification. The main clustering techniques are presented in [157, 55, 52, 93] and will be shortly reviewed in the keynotes (Sect.A.2.) Dubes [54] and Milligan and Cooper [121] proposed solutions to the choice of the number of clusters, which are related to the stopping rule in hierarchical methods. Bock [22] and Gordon [69, 70] are particularly interested in the validity assessment. Their approach is close to an *a contrario method*: They define a background model in which they measure the concentration of points. A uniform model may not be the most adapted method, and it may be useful to define a data-dependent background model as shall be done in the next chapter. The method of the present chapter is directly inspired by Desolneux *et al.*'s method for detecting dots in an image [50]. In this method, a hierarchical classification of

the set of dots is considered, and meaningful clusters are detected *a contrario* to a standard Poisson null model. A maximality criterion was also defined but had several flaws that were taken in consideration in the approach proposed in this chapter. A very complete study of hierarchical segmentation and representation is presented by L. Guigues in his PhD thesis [74] (in French).

Chapter 9

Grouping spatially coherent meaningful matches

his chapter intends to form coherent groups between matching shape elements into a shape. Each pair of matching shape elements leads to a unique transformation (similarity or affine map.) A natural way to group these shape elements into shapes is to find clusters in the transformation space. The theory developed in the previous chapter is immediately applicable. The main problem addressed here is the correct definition and computation of the *background model* π. This background model is a probability distribution on the set of similarities, or on the set of affine transforms. In order to have accurate shape clusters, π must be built from empirical measurements on the observable shape matching transforms. As in Chap. 6, the main issue is to compute accurately a density function in high dimension (4 or 6) with relatively few samples. The found solution is analogous: figure out the marginal variables for which an independenceassumption is sound. Then the density functions of these marginal laws can be accurately estimated on the data and yield an accurate *background model*.

9.1 Why a spatial coherence detection?

Looking at Figure 9.1, everybody can obviously recognize on the bottom left image a detail of Picasso's painting *Guernica* shown on the top left image. However, the painting is incomplete and partially occluded in the bottom image. It is also deformed by the perspective view. Moreover, the compression rates are also different. Fig. 9.2 displays the shape elements common to these two images, both local and global, with an affine invariant encoding. It turns out that local shape elements are much more discriminative. Indeed, since no restriction is made on the affine distortion, a lot of normalized convex shape elements look quite the same. The matching pairs have been computed by the method of Chap. 6. There are 94, whereas more global matches are due to quasi convex shape elements.

The objective of this chapter is twofold: first, to prove that shape elements corresponding to a single shape can be accurately grouped together. Second, that this grouping procedure is robust enough to discard all false matches. Incidentally, this will dramatically reinforce the confidence in the previous detections. The group NFA's are indeed usually very small.

The plan of this chapter is as follows. In Section 9.2, the parameterization of similarities or general affine transformations is described. Section 9.3 applies the general clustering ideas presented in Chap. 8, first by defining a dissimilarity measure between transformations, then by defining a suitable background model on the sets of transformations. A few experiments are also shown to illustrate the ideas. Many more results will be given in the next chapter.

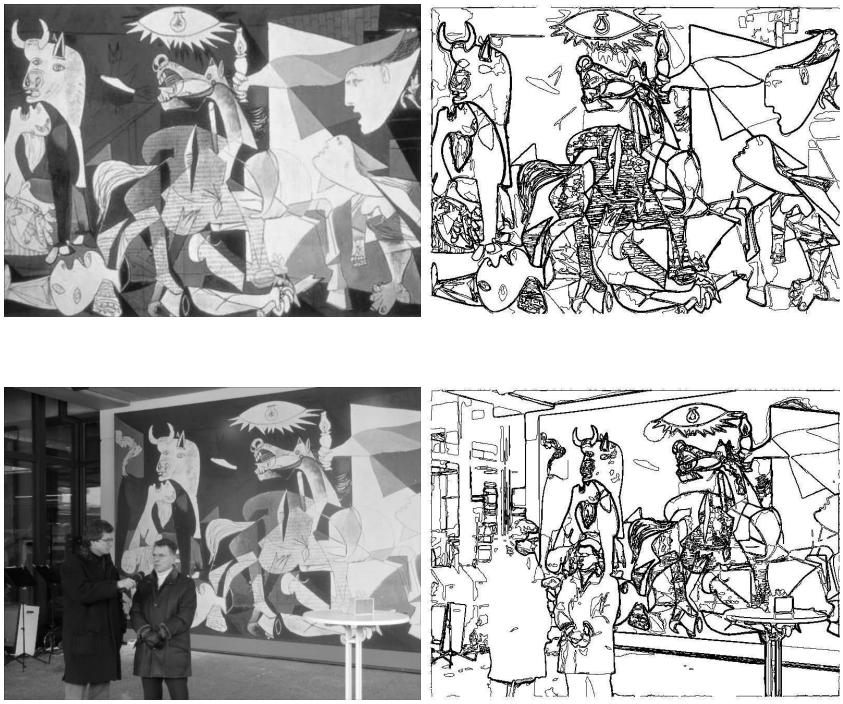


Figure 9.1: "Guernica" experiment. Original images and maximal meaningful level lines. Top: target image, bottom: scene image

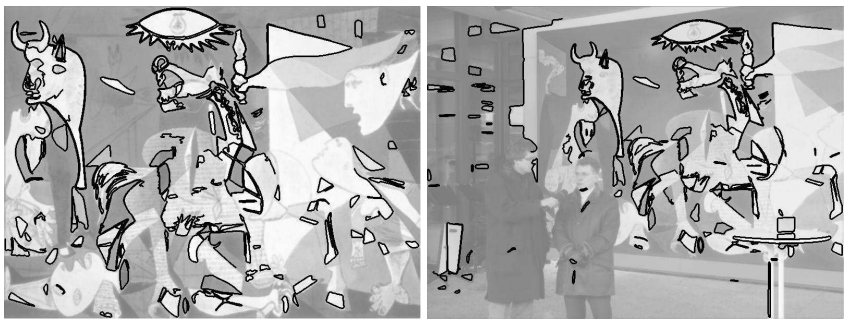


Figure 9.2: "Guernica" experiment: meaningful matches, both affine invariant semi-local and global encoding. The number of semi-local shape elements is 7440 in the first image and 6131 in the second one (hence $4.6\,10^7$ tests). The number of globally encoded shape elements is 740 (resp. 897). There are very few false matchings for locally encoded shapes and their NFA is more than 0.4. The total number of local (resp. global) matches is 94. (resp. 337.) Globally encoded shapes yields many affine invariant matches because all parallelograms, triangles or ellipses are equivalent.

9.2 Describing transformations

Let $\mathcal{I}$ and $\mathcal{I}'$ be two images, referred to as the *target* image and the *scene* image. For each match between a shape element $\mathcal{S}$ in $\mathcal{I}$ and a shape element $\mathcal{S}'$ in $\mathcal{I}'$, a geometric transformation (a similarity or an affine transform) can be computed. In what follows, the parameters involved in these transformations are described, as well as the way they can be estimated, both for the similarity and the affine transformation cases.

9.2.1 The similarity case

Let $\mathcal{S}$ and $\mathcal{S}'$ be two matching shape elements. Recall that a shape element is a normalized piece of level line described in a local frame. It is completely determined by two points, or equivalently a point and a vector. This last representation will be chosen. A local frame is then given by a couple (p, v) where p gives the origin of the frame and v gives its scale and orientation. Let us assume that $\mathcal{S}$ is related to (p, v) and $\mathcal{S}'$ to (p', v'). Since $\mathcal{S}$ and $\mathcal{S}'$ match, they differ by a similarity transformation. Now, there exists a unique similarity mapping the local frame (p, v) onto (p', v'). (See Figure 9.3.) By using complex numbers notations, this similarity can be uniquely expressed as

$$\forall z \in \mathbb{C},\ \mathbf{T}(z) = az + b,\ \text{with } a = \frac{v'}{v} \text{ and } b = p' - ap, \tag{9.1}$$

with $(a, b) \in \mathbb{C}^2$.

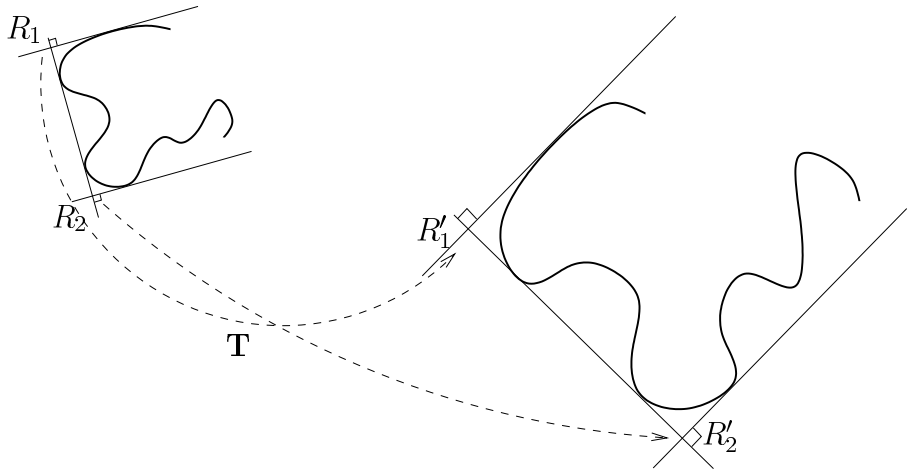


Figure 9.3: Two pieces of level lines and their corresponding local similarity frames. The similarity $\mathbf{T}$ maps R_1 into R'_1 and R_2 into R'_2

9.2.2 The affine transformation case

Let us now consider the local affine invariant normalization described in Chap. 5. Affine normalization of a piece of curve was performed by mapping its local frame $\{R_1, R_2, R_3\}$ onto the triplet $\{(0,0), (1,0), (0,1)\}$. Given another triplet $\{R'_1, R'_2, R'_3\}$ of non aligned points, there is a unique affine transform mapping $\{R_1, R_2, R_3\}$ on $\{R'_1, R'_2, R'_3\}$, again denoted by $\mathbf{T}$. There exists a unique 2×2 matrix $\mathbf{M}$ and a unique $(t_x, t_y) \in \mathbb{R}^2$ such that

$$\mathbf{T}(x, y) = \mathbf{M} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

Calculating $\mathbf{M}$ boils down to the solution of a 2×2 linear system. By the classical QR decomposition [68], $\mathbf{M}$ can be written

$$\mathbf{M} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1 & \varphi \\ 0 & 1 \end{pmatrix} \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}. \tag{9.2}$$

This decomposition is unique and completely determines $(\theta, \varphi, s_x, s_y)$ in $[0, 2\pi) \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+$. Let us denote by (x_{R_1}, y_{R_1}) and by (x'_{R_1}, y'_{R_1}) the pair of coordinates of R_1 and R'_1 respectively and by (m_{ij}) the coefficients of $\mathbf{M}$. Then, the transformation parameters $T = (\theta, \varphi, s_x, s_y, t_x, t_y)$ can be computed by means of

the following formulas

$$\begin{aligned}
\theta &= \arctan(m_{21}/m_{22}),\\
\varphi &= (m_{11}m_{12} + m_{21}m_{22}) \,/\, (m_{11}m_{22} - m_{12}m_{21}) \,,\\
s_x &= \sqrt{m_{11}^2 + m_{21}^2},\\
s_y &= (m_{11}m_{22} - m_{12}m_{21}) \,/\, \sqrt{m_{11}^2 + m_{21}^2},\\
\begin{pmatrix} t_x \\ t_y \end{pmatrix} &= \begin{pmatrix} x'_{R_1} \\ y'_{R_1} \end{pmatrix} - M \begin{pmatrix} x_{R_1} \\ y_{R_1} \end{pmatrix}.
\end{aligned} \tag{9.3}$$

Again, the vector T characterizes the transformation $\mathbf{T}$.

Without risk of ambiguity, one can adopt the same notation for similarities or affine transformations. In addition, since T characterizes $\mathbf{T}$, both of them can be identified. Thus write, for $X \in \mathbb{R}^2$, $T(X)$ instead of $\mathbf{T}(X)$.

Figure 9.4 shows the three 2-D projections of the transformation points T_k corresponding to the "Guernica" affine invariant meaningful matches (Figure 9.2).

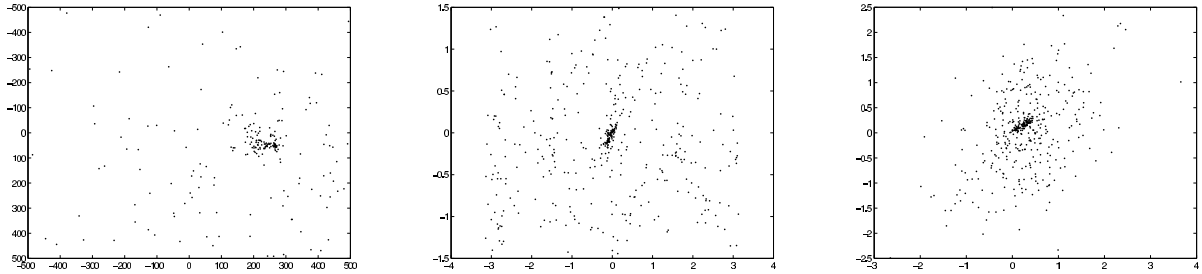


Figure 9.4: "Guernica" experiment: Each point represents a transformation associated with an affine invariant meaningful match, described by 6 parameters. Each figure represents a two-dimenstional projection of the points, respectively t_x vs. t_y (translation coordinates), θ (rotation) vs. φ (shear), and $\ln(s_x)$ vs. $\ln(s_y)$ (zooms in the x and y) directions. The noise is mainly due to similar global shape elements, which do not belong to the same real shape. The main cluster is also spread because of the effect of perpective.

9.3 Meaningful clusters of transformations

The problem of planar shape detection is by now reduced to a clustering problem in the transformation space. According to Chap. 8, it is necessary to define

1. a dissimilarity measure between points in the transformation space,
2. a probability on the space of transformations,
3. a grouping strategy.

9.3.1 A dissimilarity measure between transformations

Defining a distance between transformations is not trivial, for two reasons. First, the magnitudes of the parameters of a transformation are not directly comparable. This problem is not specific to transformation clustering

but general to clustering of any kind of data, and is discussed in Sect.A.2. Second, our representation of similarities or affine transformations does not behave well in a vector space. A sound distance is not necessarily derived from a norm.

DEFINITION 9.1 (SIMILARITY CASE) *Let T (resp. T') be the similarity determined by two shapes elements $(\mathcal{S}_1, \mathcal{S}_2)$ (resp. $(\mathcal{S}'_1, \mathcal{S}'_2)$). Let also (R_1, R_2) (resp. (R'_1, R'_2)) be the points determining the local frame of $\mathcal{S}_1$ (resp. $\mathcal{S}'_1$). We call dissimilarity measure of T, T',*

$$d_S(T, T') = \max\left\{\|T(R_i) - T'(R_i)\|,\ \|T(R'_i) - T'(R'_i)\|,\ i \in \{1, 2\}\right\}. \tag{9.4}$$

LEMMA 9.1 *The function d_S is non negative, symmetric and satisfies $d_S(T, T') = 0 \Leftrightarrow T = T'$.*

Proof: The first two properties are obvious. Since a similarity is uniquely defined by the images of two points, the last property follows. Remark that d_S is not a distance since the triangle inequality does not hold. □ For a sake of completeness, let us define a dissimilarity between affine transforms.

DEFINITION 9.2 (AFFINE CASE) *Let T (resp. T') be an affine transform determined by two shapes elements $(\mathcal{S}_1, \mathcal{S}_2)$ (resp. $(\mathcal{S}'_1, \mathcal{S}'_2)$). Let also (R_1, R_2, R_3) (resp. (R'_1, R'_2, R'_3)) the points determining the local frame of $\mathcal{S}_1$ (resp. $\mathcal{S}'_1$). We set*

$$d_A(T, T') = \max\left\{\|T(R_i) - T'(R_i)\|,\ \|T(R'_i) - T'(R'_i)\|,\ i \in \{1, 2, 3\}\right\}. \tag{9.5}$$

9.3.2 Background model: the similarity case

In order to apply the detection framework of Chap. 8, a background law is first needed. A data point here is a similarity transformation represented by a pair of complex numbers $(a, b) \in \mathbb{C}^2$. The purpose of this section is to devise a sound background law π on the set of similarity transformations. To this aim, recall that (a, b) is determined by two local frames in the images to be matched, respectively (p, v) and (p', v'). Let us now assume that these observations are the realization of a random variable $(P, V, P', V') \in \mathbb{C}^4$. It is natural to assume that the position, the size and the orientation of an object are independent. This is certainly sound, up to some border effects. In addition, two images which do not contain common shapes also can be assumed independent. This leads us to take to the following independence assumption for the background model.

(A') *Consider a random model image $\mathcal{I}$ and a random scene image $\mathcal{I}'$. Then the random variables P, $|V|$, $\arg V$, P', $|V'|$, $\arg V'$ associated with (necessary casual) matches between both images are mutually independent.*

The marginal laws of the six previous random variables can easily be learned from the two images. Hence, the law of (P, V, P', V') is assumed to be known. By (9.1), such a 4-tuple uniquely defines a random similarity pattern denoted by (A, B), where A represents the rotation and zoom, and B the translation. The background law π is nothing but the distribution of (A, B). The expression of (A, B) as a function of (P, V, P', V') is explicit and given by

$$(A, B) : (P, V, P', V') \mapsto \left(\frac{V'}{V}, P' - \frac{V'}{V}P\right).$$

The background law π is the image of the law (P, V, P', V') by this application. It is also clear that A and B are not independent. Nevertheless, by definition of the conditional law,

$$d\pi(a, b) = d\pi^B(b \,|A = a)\, d\pi^A(a), \tag{9.6}$$

where π^A is the marginal of A and $\pi^B(\cdot\,|A = a)$ is the law of B knowing $A = a$. Since $|A| = |V'|/|V|$ and $\arg A = \arg V' - \arg V \mod (2\pi)$, these two variables are independent under Assumption **(A')**. Thus, the distribution π^A can easily be computed. Moreover, it turns out that A is independent from P and P'. Hence, the law of $B = P' - AP$, conditionally to $A = a$ is the law of $P' - aP$, which can also be easily computed under **(A')**. The background law π follows from (9.6).

In practice, the computation of π between two images is as follows:

1. Compute all the shape elements of model and target images.
2. Compute the empirical laws of P, V, P', V' giving the position, the scale and the orientation of the local frames related to shape elements in the two images. Under the independence assumption **(A')**, this yields the law of the background model (P, V, P', V').
3. Under the same assumption, compute the empirical laws of $|A| = \frac{|V'|}{|V|}$ and $\arg A = \arg V' - \arg V \mod (2\pi)$.
4. For each value a of A with non null frequency, compute the empirical distribution of $P' - aP$.

The probability of a region R is then given by approximating the integral

$$\pi(R) = \int_R d\pi^B(b|A = a)\, d\pi^A(a).$$

A few words about the estimation of the background model. One would expect $\arg A$ to be uniformly distributed in $[-\pi, \pi)$, and this belief was experimentally confirmed. We refer to Figure 9.5(a) for an example of empirical distribution of $\arg A$ from the "Guernica" experiment of Figures 9.1 and 9.2. The distribution of the zoom factor $|A|$ is instead far from being uniform. Figure 9.5b shows an example of the empirical distribution of $\ln(|A|)$ from the same "Guernica" experiment. There is no way to figure out a realistic *a priori* distribution for $|A|$, or for B given A. The background model distributions must be learned from the scene and target images.

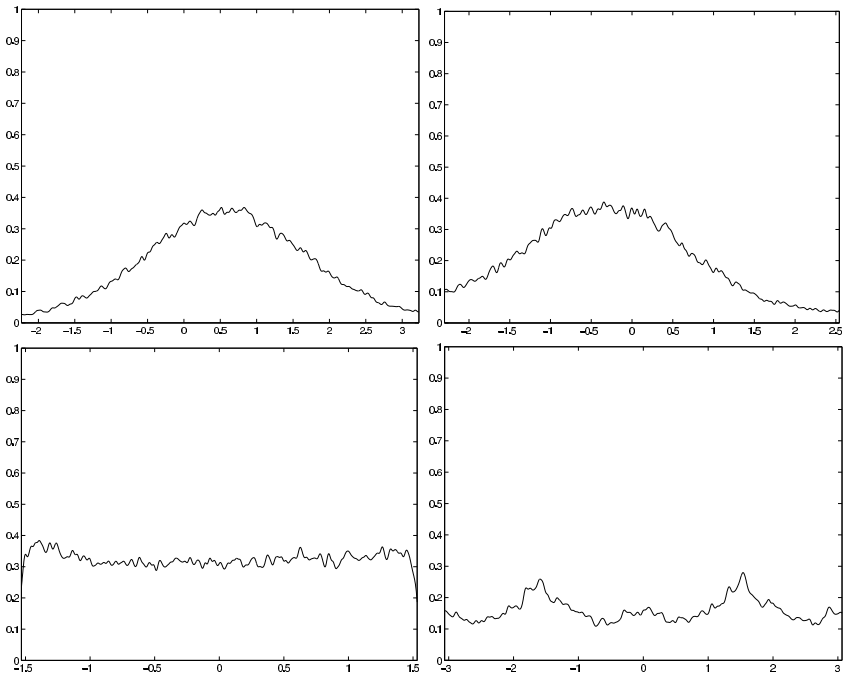


Figure 9.5: Empirical histograms for affine invariant matching for the experiment of Fig. 9.1. On the first row, the empirical zoom factors in the x and y direction (logscale), which are image dependent. On the second row, the distribution of the shear and the rotation angle. The shear is basically uniform, but the rotation exhibits some slight peaks around $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ because of the numerous lines in the image.

Remark: The ideas presented here also hold for the affine transformation clustering. For this case, θ, φ, s_x and s_y are considered to be mutually independent. Their distributions can be learnt empirically, as well as the joint probability of (t_x, t_y) given $(\theta, \varphi, s_x, s_y)$. This construction, experimentally satisfying though it is (see next chapter), has no righteous theoretical justification. The problem of finding the right independent marginal variables in the affine case is left open.

9.4 Experiments

The consistency of the previous definitions is now briefly empirically checked. The next chapter will contain many more experiments. All the experiments will be performed with a pair of images. It is worth summarizing the steps leading to a complete experimental setting for shape identification.

1. The method of Chap. 6 is first applied and yields a set of M pairs of matching shape elements, one in the target image and one in the scene image.

2. A background model π on the set of similarities or on the set of affine transforms E is built according to the method of the present chapter.

3. The transforms $T_1, \ldots, T_M$ associated with the matching pairs form a point data set in E. From this set, a clustering tree is built according the dissimilarity measures of Definitions 9.1 or 9.2 (affine case.)

4. Maximal groups are computed by Def. 8.5.

The final outcome of the shape identification method of this book is, for each pair of images, a set of maximal meaningful clusters. Each cluster is likely to correspond to an identified shape. One can display for each cluster its associated shape elements. If the grouping is correct, this set of shape elements must correspond to a *matching shape* in both the target image and the scene image. In practice, the identified shapes have dramatically low NFA's. Thus, they yield an overwhelming certainty about identification. This certainty is, however, not fully unambiguous because of the Strobe effect. Indeed, shapes often have self-similar parts: windows, or rows of windows in a building are a good example. Other examples are given by symmetries. For instance, the letter N is self-similar by a π rotation. In cases, two or more very meaningful groups can be found, each one corresponding to a shape self-similarity. Such self-similarities can, however, easily be anticipated by a previous comparison of the target image with itself. This comparison can be performed by the above algorithm. The main group with then correspond to the global match of the shape with itself and the other groups to Strobe effects.

Figure 9.6 depicts the maximal meaningful groups for the "Guernica" experiments. There is one single maximal meaningful group, with $-\log_{10}(NFA_g) = 196.2$. The best match between shape elements has a NFA about $4.16\ 10^{-12}$. Hence grouping dramatically increases confidence in detections, while all the false matches are eliminated.

Figure 9.6: "Guernica" experiment: a single maximal meaningful group was detected. Matches of the group for the target image (left) and the scene image (right). The group is composed by 117 good matches, and its $-\log_{10}(NFA_g)$ is 196.2.

In Fig. 9.9 ("Casablanca" experiment, see the original images in Fig.7.14), two maximal meaningful groups are detected.

The indivisibility criterion (8.12) decides that two separate groups (the actors' faces on the one hand and the word "Casablanca" on the other hand) are a better representation than a single large group containing both

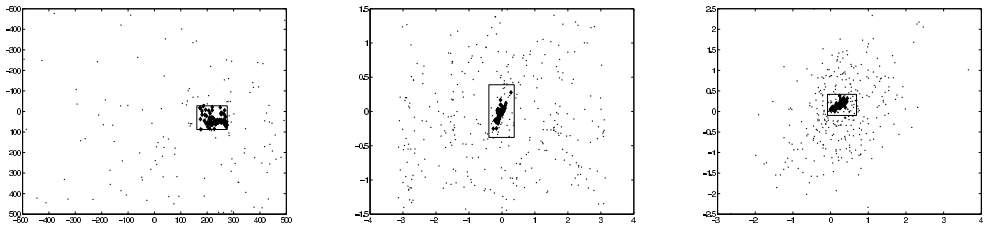


Figure 9.7: "Guernica" experiment: data points of Fig. 9.4. The plotted rectangle is the region attaining the NFA of the only meaningful group. All the other points are considered as isolated and do not belong to any group.

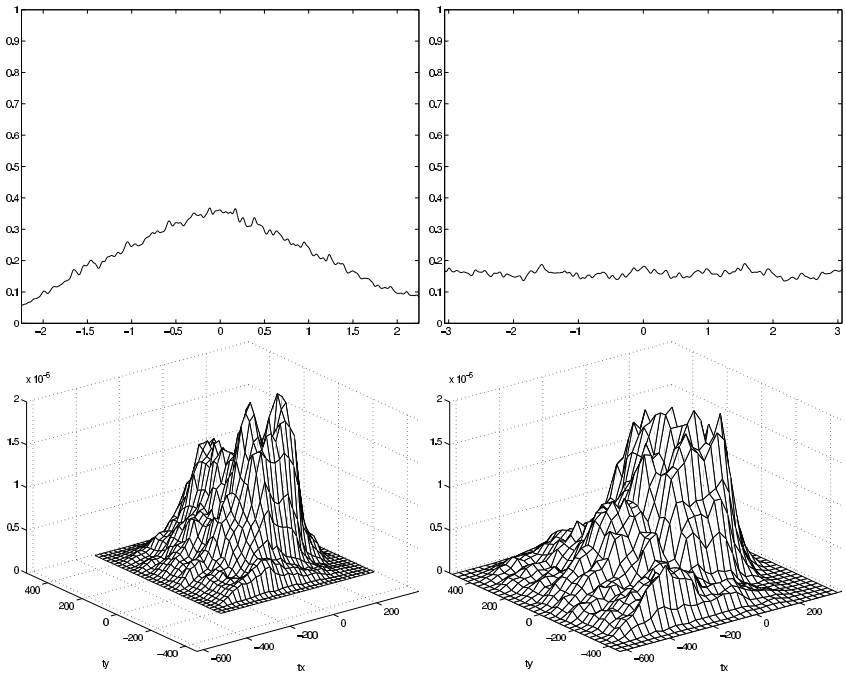
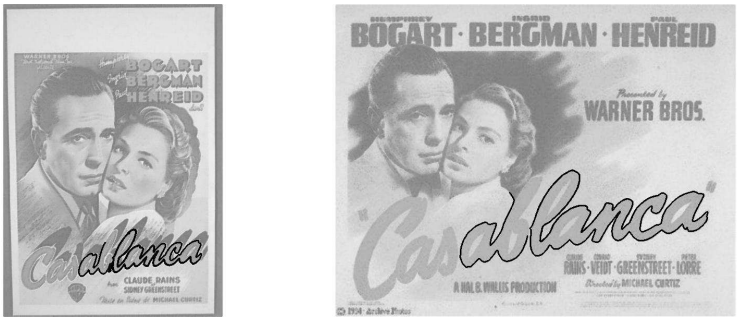
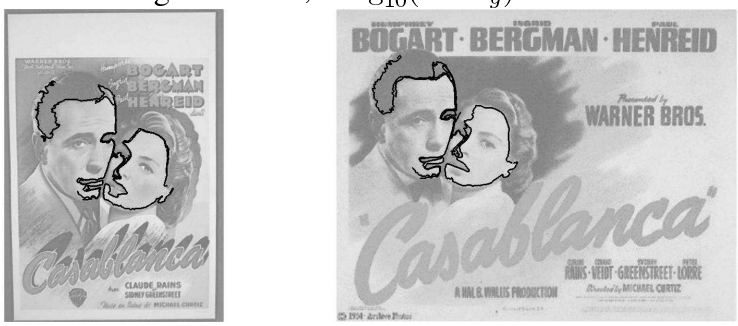


Figure 9.8: Empirical histograms for similarity invariant matching for the experiment of Fig. 9.9. On the first row, the log-empirical zoom factor $\ln(s)$ and the rotation angle θ. This last one is nearly uniform in this case. On the bottom row, the distribution of the translation vector, conditioned by two different values of the couple $(\ln(s), \theta)$. These values correspond to the two maximal groups that are depicted on Fig. 9.9. Since the scales are different, so are the distributions.

groups. Indeed, while the large group in Figure 9.10 has a lower NFA_g than one of its children ($10^{-31.9}$), it is not indivisible. Indeed, the NFA_g of its two children are $10^{-32.85}$ and $10^{-17.62}$. By Prop.8.4, the largest group is not indivisible, and thus cannot be maximal.



12 meaningful matches, $-\log_{10}(NFA_g) = 32.85$



8 meaningful matches, $-\log_{10}(NFA_g) = 17.62$

Figure 9.9: "Casablanca" experiment: maximal meaningful groups.

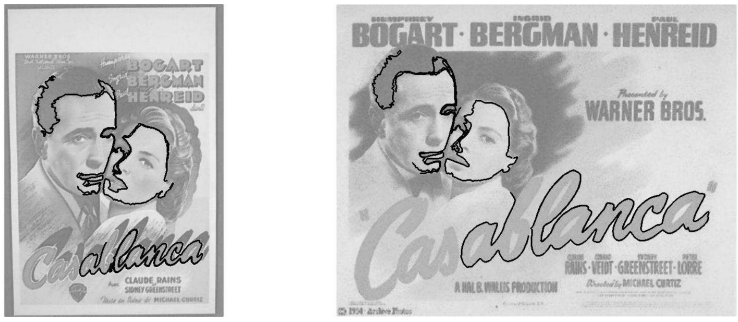


Figure 9.10: "Casablanca" experiment. Meaningful group corresponding to the merging of groups in Figure 9.9. This group contains 20 meaningful matches, and its $-\log_{10}(NFA_g)$ is 31.9. According to Definition 8.4 and following Prop. 8.4, it is not indivisible and cannot be maximal.

9.5 Bibliographical notes

The use of spatial coherence for shape or object detection has been the subject of intensive research, in particular since Ballard's work on the generalized Hough transform [14]. In his paper, Ballard proposed a method extending the Hough transform to any kind of planar shape, not necessarily described by an analytic formula. Stockman [156] presented another early work based on the same principle (recognize a target shape by finding clusters in the transformation space), where he introduced a coarse to fine technique allowing to reduce the search complexity. Other voting schemes, like Geometric Hashing [170, 100] or the Alignment method [85], are frequently used in the detection or recognition problem. They are computationally more expensive and can be less accurate. In [72, 73] Grimson and Huttenlocher presented a study on the likelihood of false peaks in the Hough parameter space. Their work inspired the detection method adopted in this chapter. They indeed proposed a detection framework where recognition thresholds are derived from a null model (*"the conspiracy of random"*). Previous recognition methods generally associated a single threshold with each target image, independent of the scene complexity. In contrast to these methods, the grouping thresholds derived in this chapter satisfy an important property: they are functions of the scene complexity and of the uncertainty in feature extraction. The method of the present chapter took these fundamental ideas from Grimson and Huttenlocher's work. The computational swiftness is obtained by a hierarchical representation of the transformation points. The definition of a data-dependent background model is crucial for avoiding false clusters: Grimson and Huttenlocher's method assumes that matched features are uniformly distributed in the image. This assumption is usually not valid; see [139].

Finding groups in data sets is a major problem in many fields of knowledge such as statistical pattern recognition, image processing, or data mining. Grouping phenomena are probably essential in human perception. In vision, the grouping phenomenon was thoroughly explored by the Gestalt school. The founding paper on this problem definitely is Wertheimer [165] . In Computer Vision, the first attempts to model a computational perceptual organization date back to Marr [114]. More recently D. Lowe [111] proposed a detection framework based on the computation of accidental occurrences. He writes:

> *In other words, one can shift our attention from finding properties with high prior expectations to those that are sufficiently constrained to be detectable among a realistic distribution of accidentals.[...] Even when we ignore the ultimate interpretation for some grouping and therefore its particular* a priori *expectation, we can judge it to be significant based on the non-accidentalness criteria.*

In the beautiful paper [104], M. Lindenbaum proposed to evaluate *a priori* the performance of any invariant shape recognition device. For this author, a shape can be distinguished if and only it could occur with very small probability in the random background. The author gives a lower bound on the number of points k in the shape ensuring recognition. This lower bound depends upon: the number of points N in the background, the accuracy d of the recognition, the required invariance and ε, the allowed for error probability for each test. Unfortunately, the shape model assumed by the author is not quite realistic. For him, a shape is a cluster of points concentrated around some curve representing the shape's boundary and the background is modelled as a Poisson noise with lower density.

In [21], the authors have proposed a probabilistic *compositional model* for shape recognition. In their vision model, visual primitives are recursively composed, subject to syntactic restrictions, to form tree-structured objects. The involved compositional rules have a structure close to Chomsky's grammars. To take a trivial but significant example, an example of syntactic rule may be :

$$alignment + alignment \rightarrow alignment,$$

with the restriction that both alignments are themselves aligned. From the probabilistic viewpoint, the source of inspiration of this theory is very close to our aims. This is best illustrated by the following quotation of Laplace's Essay on Probability which we take from [21]. *"On a table we see letters arranged in this order,* Constantinople*, and we judge that this arrangement is not the result of chance, not because it is less possible*

than the others, for if this word were not employed in any language we should not suspect it came from any particular cause, but this word being in use amongst us, it is incomparably more probable that some person has thus arranged the aforesaid letters than this arrangement is due to chance". Laplace is assuming *a contrario* that any combination of the 26 alphabet letters would be equally likely. Now, a modern dictionary contains not more than 10^5 words. The number of possible words with 14 letters like Constantinople in the *a contrario* model is about $2.4\,10^{19}$. Thus the probability of the group Constantinople happening “just by chance" is less than 10^{-14} in the *a contrario* model.

Chapter 10

Experimental results

he grouping of spatially coherent meaningful matches has been extensively studied in the previous chapter, and several experiments were presented and discussed. In this chapter we illustrate the whole recognition process by presenting some more experiments over different kinds of images.

10.1 The visualization of the results

Almost all the experiments presented in this chapter are illustrated with the following images:

1. *The two original images.*

2. *The smoothed maximal meaningful boundaries of the original images*, extracted using the algorithm described in Chapter 3, then smoothed with Moisan's implementation of the affine curve shortening equation (Chapter 4.3).

3. *Detection of meaningful matches between shape elements.* We consider here the 1-meaningful matches, despite the fact that a few of them may correspond to false detections; indeed, as seen in Chapter 6, the constraints imposed by the encoding methods and by the non-intersection of level lines introduce a certain amount of dependence between the distances used as features in the *background model* (which were assumed to be independent). Thresholding the NFA at 0.1 empirically ensures that no detection can occur in white noise images. However, since the detection of meaningful matches is followed by a grouping process based on spatial coherence, in the experiments these few false matches are kept in order to test the robustness of the grouping algorithm.

   A fundamental hypothesis for the *a contrario* detection of groups is that, under the *background model*, transformation points are mutually independent. In order to comply with this hypothesis, a greedy algorithm that eliminates matched shape elements which share a large piece of curve with other shape elements presenting lower NFA. More precisely speaking, if a pair of shape elements $(\mathcal{S}_1, \mathcal{S}'_1)$ is an ε_1-meaningful match, and there exists another pair $(\mathcal{S}_2, \mathcal{S}'_2)$ matching ε_2-meaningfully, with $\varepsilon_2 < \varepsilon_1$, such that $\mathcal{S}_1$ shares at least half of its length with $\mathcal{S}_2$, and the same for $\mathcal{S}'_1$ and $\mathcal{S}'_2$, the pair $(\mathcal{S}_1, \mathcal{S}'_1)$ is eliminated from the output list of matches.

   The detection of 1*-meaningful matches is illustrated by superimposing the matched shape elements to the original images.*

4. *Grouping of spatially coherent meaningful matches.* For each meaningful group of matches that is detected (the maximal 1-meaningful groups defined in Chapter 9), four images are shown:

   - *The shape elements that match within a group are shown, superimposed to the original images.*

- Given the set of transformations corresponding to the matches within a group, the best affine or even projective transformation (in the least squares sense) that maps the shape elements in the target image to the ones in the scene image is computed. Then, the target image is mapped using this transformation. *The superposition of the transformed target image and the scene image is presented.*
- All the pieces of meaningful level lines of the two registered images are then compared, as a visual check. To this purpose, let us fix two values l and d. Let C_1 and C_2 two pieces of level lines with the same length l, parameterized by the length parameter. If for all $s \in (0, l)$, $|C_1(s) - C_2(s)| < d$ then display C_1 and C_2.

10.2 Experiments

The detection framework presented in former sections is completely general and can be applied to any kind of images, provided objects are well described by planar shapes and transformations are close to similarities (or affinities). Besides the "Guernica" and "Casablanca" experiments, this section gives some examples of different kind and nature. All experiments were done using the single-linkage algorithm (see Chapter 8, Section A.2.2).

Mutiple occurrences of a logo

This example illustrates the performance of the proposed methodology in detecting multiple groups in an image. Two images contain occurrences of (parts of) the Coca-Cola logo are compared on Figure 10.1). Figure 10.2 shows the meaningful matches, both locally and globally encoded with the affine invariant method. They lead to points in the 6-dimensional space, clustered by the single linkage methods. Maximal meaningful groups appear on three projections in Fig. 10.3. The corresponding shape elements are displayed for each group in Fig. (10.4) and (10.5). Five groups are detected and are all correct. The NFA_g of maximal meaningful groups are reported in Tab. 10.1.

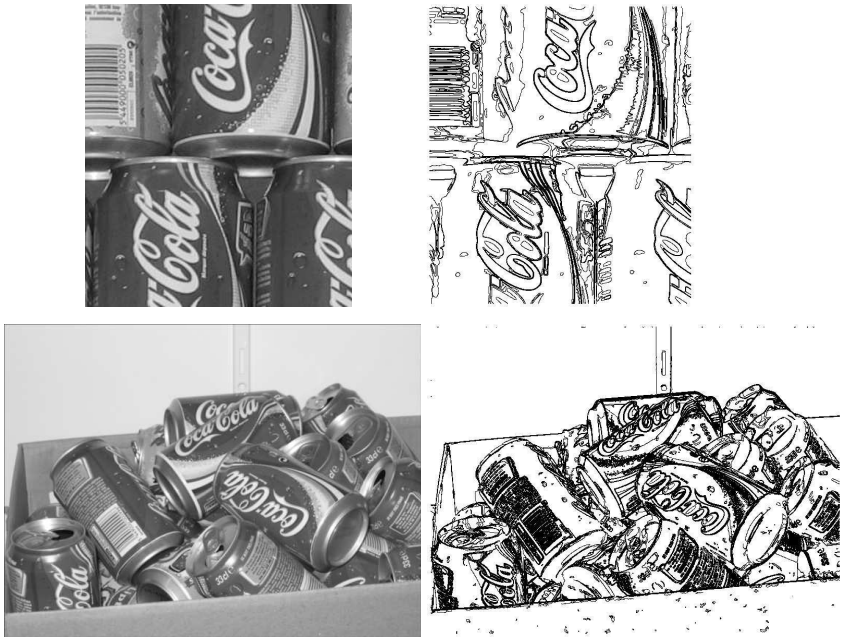


Figure 10.1: "Coca-Cola" experiment: original images and maximal meaningful level lines. Top: target image, bottom: scene image.

Figure 10.2: "Coca-Cola" experiment: meaningful matches, with local encoding (top) and globally encoding (down). Number of tests: $1.57\ 10^7$ (590 shape elements in the target image, $26,620$ in the scene image). There are 133 meaningful local matches, 1002 global matches. The best match has $NFA = 8.4\ 10^{-12}$.

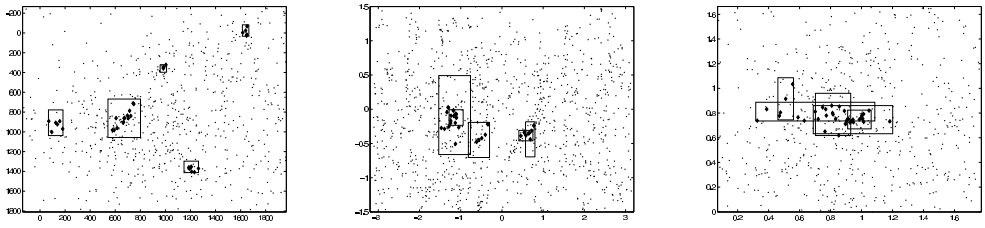


Figure 10.3: "Coca-Cola" experiment: meaningful groups and the projection of the regions. Their $-\log_{10}(NFA)$ is respectively 19.4, 13.5, 3.7, 1.9 and 0.6. The image on the top left corresponds to the projection on the (t_x, t_y) plane, where the groups are clearly separated. The second plot displays the rotation θ against φ (the shear). Finally the last figure depicts the zoom in the x and y coordinates in the normalized frame (logarithmic scale). Finding those groups from the clouds does not seem trivial.

Table 10.1: "Coca-Cola" experiment: NFA_g for the maximal meaningful groups in Figure 10.4

Group nb.	1	2	3	4	5
nb. of matches	15	7	5	6	4
$-\log_{10}(NFA_g)$	19.4	13.5	3.7	1.9	0.6

Église de Valbonne

Figure 10.8 shows two different views of the *Église de Valbonne*, with their corresponding maximal meaningful level lines. The meaningful matches between these two views, up to similarity invariance, are shown in Figure 10.9. Some of them are false matches, but all of them showed a NFA greater than 0.1, as predicted by the experimental results in Chapter 6. There are also some casual matches that correspond to the same structures in the images. Figure 10.2 displays the only detected maximal meaningful group (see caption for details). A global affine transformation was estimated from this group by means of a least squares procedure, over the cor-

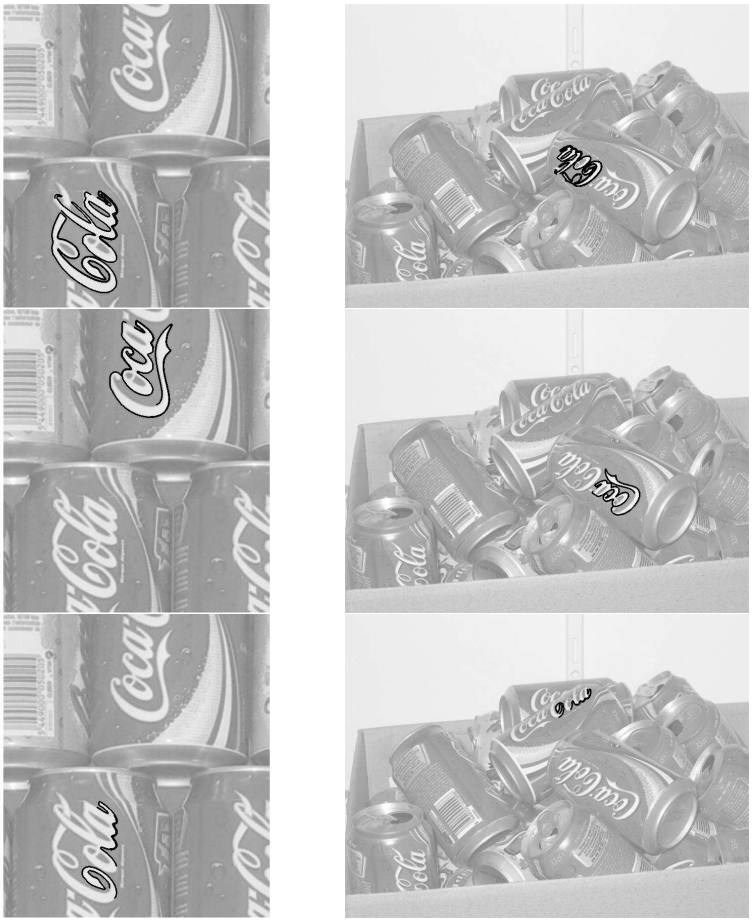


Figure 10.4: "Coca-Cola" experiment: first three maximal meaningful groups (among 5). Their $-\log_{10}(NFA)$ is respectively 19.4, 13.5, 3.7.

Figure 10.5: "Coca-Cola" experiment: maximal meaningful groups (last two among five). Their $-\log_{10}(NFA)$ is respectively 1.9 and 0.6.

Figure 10.6: "Coca-Cola" experiment: superposition of the logo onto the image, for the first three groups. A mean transformation is computed for each group of transformations by using a linear regression. Another practical way to check the validity of the transformation is to display all the pieces of maximum meaningful level lines that are everywhere close to each other after registration (in practice pieces of length 40 at distance less than 4).

Figure 10.7: "Coca-Cola" experiment: superposition of the logo onto the image, for the last two groups. See caption of Fig. 10.6 above.

responding matched shape elements. These transformations were used to map the target image into the scene image (Figure 10.11). The superimpositon of the transformed target image and the scene image shows that the estimated transformation is a good approximation.

Figure 10.8: Two frames of the *Église de Valbonne* sequence, with its corresponding meaningful level lines. The image on the top was considered as target

Figure 10.9: Église de Valbonne: 81 meaningful matches were found, for 14,710 shape elements in the target image and 18,413 in the scene image. All false detections have NFA larger than 0.1. The best match has $NFA = 2.98\ 10^{-12}$.

Figure 10.10: Église de Valbonne: a single maximal meaningful group is detected. All false matches and spatially uncoherent matches are rejected. The group contains 35 (similarity invariant) meaningful matches and $-\log_{10}(NFA_g) = 84.8$.

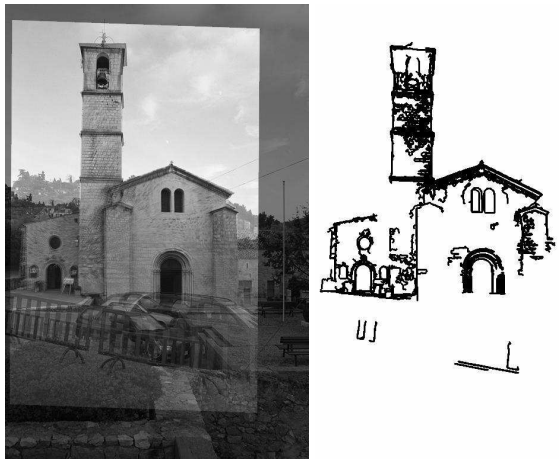


Figure 10.11: Église de Valbonne: registration of the images computed from the maximal meaningful group. On the right, the matching pieces of level lines. On the bottom right image, some straight lines appear, because of the coincidental superposition of the lower and upper part of the gate after registration.

The next experiment shows that the grouping procedure naturally induces a background/foreground separation. The two images to be matched are frames of a movie.

Figure 10.12: Tramway images. Two images of a movie.

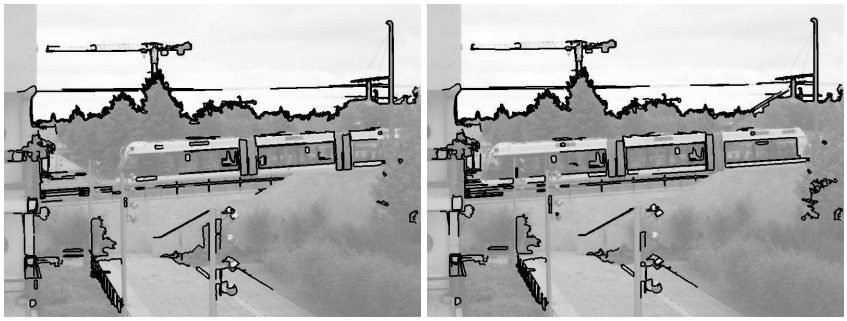


Figure 10.13: Tramway images. Meaningful matches. 434 (resp. 71) local (resp. global) shape elements match.

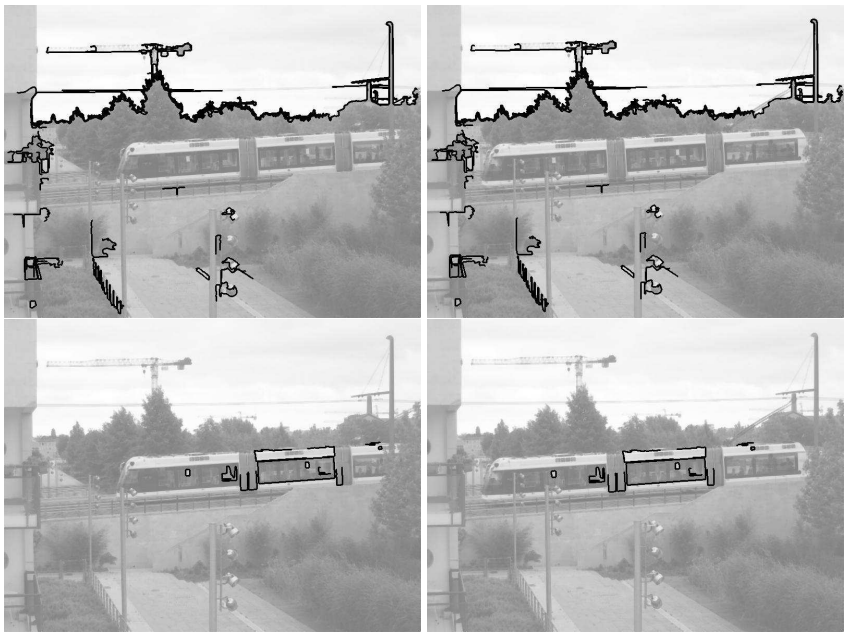


Figure 10.14: Tramway images. The two maximal meaningful groups. The first one corresponds to the background, contains 220 matches, and $-\log_{10}(NFA_g) = 643.7$. The second one is the train, with 15 matches and $-\log_{10}(NFA_g) = 11.9$.

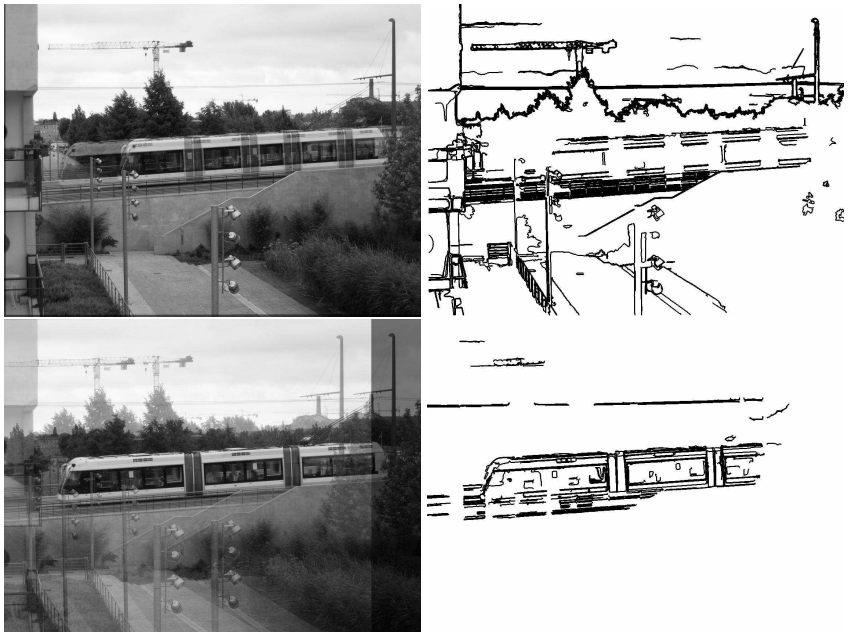


Figure 10.15: Tramway images. Registration with respect to two maximal meaningful groups. The first set of lines corresponds to the background. It is worth noticing the outer contour of the train. This is a consequence of the well-known aperture problem in optical flow computation: the visual motion cannot be determined in the direction of the level lines, since this does not result in any change in the image. The counterpart is visible on the bottom images, where the motion of (static) cables cannot be separated from the motion of the tram.

10.3 Dealing with occlusions

This section describes an example where the region of interest in the scene is occluded by the foreground. The images are two photographs of the painting *Las Meninas* by Velázquez. One is taken from the web, and the other was shot directly in the Museum of the Prado in Madrid. As can be seen in Fig. 10.16, one image is occluded by some people contemplating the painting and the colors and the illumination are completely different. Nonetheless, maximal meaningful level lines are quite insensitive to this change. This empirical statement will be proven by the matching and the grouping phase. The image on the top is considered as target image, and its shape contents is sought in the bottom image. In this experiment, the similarity version of the recognition method was used. Figure 10.17 shows the detected 1-meaningful matches between shape elements.

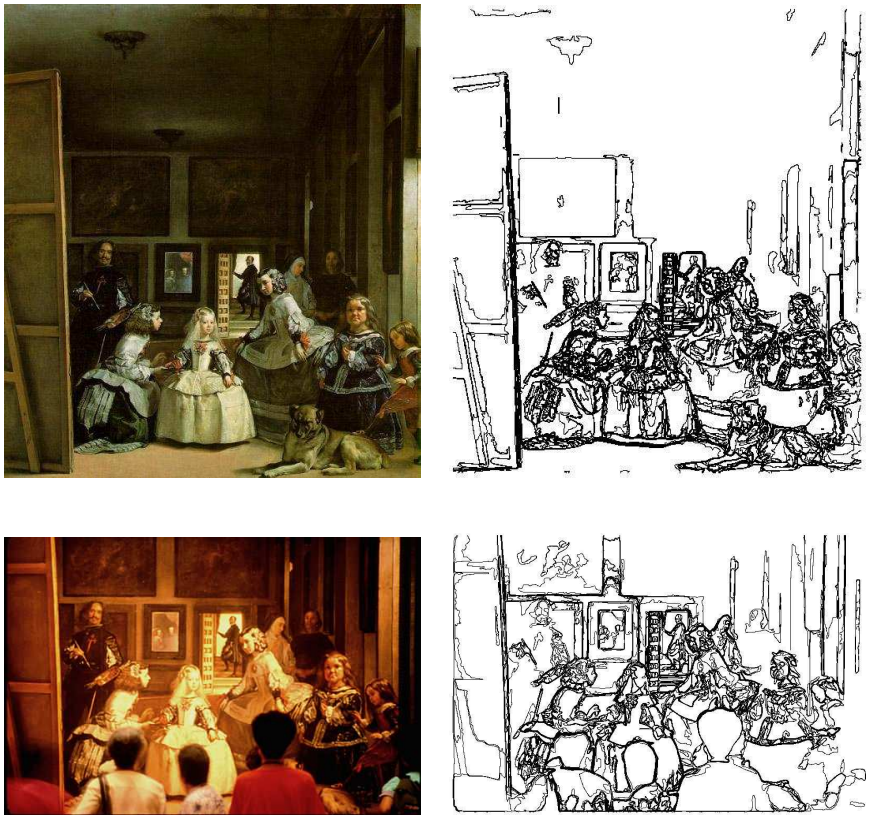


Figure 10.16: Las Meninas experiment. Top row: target image and its maximal meaningful boundaries. Bottom: scene image and maximal meaningful boundaries

The best match shows an NFA of $4.1\,10^{-14}$. Here again, the few false matches that were found have their NFA above 0.1.

A single maximal meaningful group was detected. This group contains 70 spatially coherent meaningful matches, and its $-\log_{10}(NFA_g)$ is 226.70. Figure 10.18 shows the matched shape elements that are within the group.

We end up with "Las Meninas" experiment by showing, as usual, the superposition of the registrated target image and the scene image (Figure 10.19). The registration is very accurate, as can be seen in the pieces of level lines that are common to the two images. Nearly all the shape content is matched in this case.

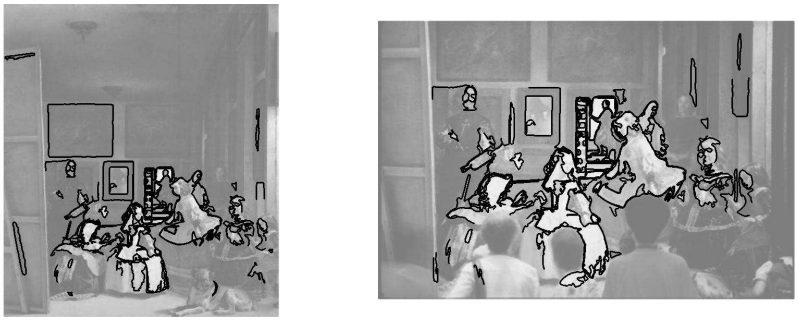


Figure 10.17: Las Meninas experiment: 1-meaningful matches. The NFA of the best match is $4.1\,10^{-14}$. Some "false" detections can be observed. All of them are due to global matches between nearly convex pieces.

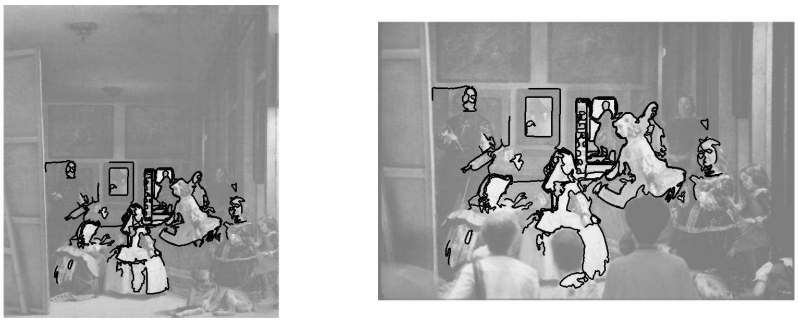


Figure 10.18: The 70 matched shape elements within the spatially coherent group. All false matches have been rejected. The value of $-\log_{10}(NFA_g)$ of the group is 226.70.

Figure 10.19: "Las Meninas" experiment. Left: superposition of the "scene image" and the transformed "target image". Right: image of pieces of level lines comparison.

10.4 Strobe effect

Strobe effect is defined as partial image autosimilarity, that is to say when parts of the same image are similar. When two images presenting Strobe effect are to be matched, then all the combination of local matches should be detected. This last example consist in finding groups of spatially coherent meaningful matches between the two images shown in Fig 10.20, that are two images of a panoramic view. Four maximal meaningful groups are detected. The most meaningful is of course overwhelming, with 169 matches and $-\log_{10}(NFA_g) = 534.8$!. The other groups are also very meaningful, though. There are completely legitimate since they correspond to repeating parts of the image. Remark that other Strobe effects should be detected.

Figure 10.20: Round about images.

Figure 10.21: Round about images. Meaningful matches. There are 274 local matches, and 645 global ones.

Table 10.2: Round about images. Number of matches and NFA_g of the meaningful groups (in the order depicted in Fig. 10.22).

Group nb.	1	2	3
nb. of matches	169	12	17
$-\log_{10}(NFA_g)$	534.8	76.81	46.2

Figure 10.22: Round about images. There are three maximal meaningful groups. The NFA_g and the number of matches are reported in Tab. 10.2. All these groups are correct, and the last three ones are due to the local autosimilarity of the image.

Figure 10.23: Round about images. The superposition of the two images when the first one is mapped onto the second one, by the affine mapping computed from maximal meaningful groups. The superposition of the two images is clear in the area of the group. It is checked on the pieces of level lines. Other vertical level lines are also visible at other places since the images contain many of them.

Appendix A

Keynotes

A.1 On the edge extraction problem

The literature on edge detection is tremendous. Hence, instead of giving a large list of references, the meaningful boundaries (MB) method will be compared with two of the most representative state of the art methods.

A.1.1 Meaningful boundaries vs. Haralick's detector

Following Haralick [76], edges are the maxima of the gradient norm in the direction of the gradient, such that the gradient is larger than a given threshold. Thus, for a grey-level image u, they are the zero-crossings of $D^2u(Du, Du)$. Since this quantity is numerically sensitive to noise, a multiscale strategy *à la* Marr is applied. In practice, u is first convolved with a Gaussian with standard deviation σ. Let us denote by g_σ this Gaussian and $u_\sigma = g_\sigma * u$. Edge pixels are defined such that $|Du_\sigma| > \mu$ and $D^2u_\sigma(Du_\sigma, Du_\sigma)$ has a different sign for neighboring pixels. There have been some attempts to automatically determine the scale parameter σ [103], but edge detection widely remains multiscale as predicted by Marr [114]. In practice, it is quite difficult to track edges back to small scales. The multiscale meaningful boundaries detection of Section 3.3 allows to consider different scales, while keeping detection thresholds completely automatic. Moreover, the number of scales has a logarithmic influence.

A second problem is that Haralick's detector provides with a set of points or curves containing only a few pixels. The way they should be connected is far from obvious. It may lead to a very high computational complexity, and depends on several sensitive parameters. Level lines are Jordan curves, and do not have this problem.

Last but not least, Haralick's operator is inefficient for corners and junctions. Indeed, at those points, the gradient direction is very badly estimated and edges may be severely cut. Additional algorithms are necessary to reconnect pieces of edges. On the opposite, level lines bifurcate at junctions, thus handling the different boundaries. (See Sect. 3.1.) Figure A.1 shows the meaningful boundaries and Canny's filter (which is an optimized version of Haralick's method) near two junctions. First, edges gives very small pieces of curves. Even though there are some linking procedures, any additional algorithm is a drawback. The behavior of the level lines around the T-junctions are quite clear. When extracting shape elements by local encoding, all the different configurations near the junctions will be considered. Clearly, meaningful level lines provide a set of curves which is more reliable and directly usable, to the expense of a more heavy computational cost (a few seconds for a typical 512×512 images, half of the time being dedicated to the computation of the tree of level lines, half to the selection of meaningful boundaries).

A.1.2 Meaningful boundaries or snakes?

Active contours is one of the most popular techniques of boundary detection. The first works of Kass, Witkin and Terzopoulos [92] have been improved and generalized by many authors. Recent models are more intrinsic,

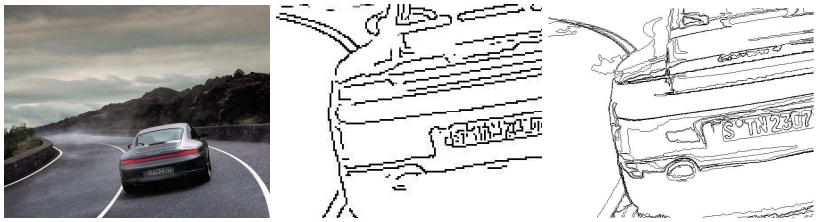


Figure A.1: Junction and level lines. On the left, the original image. Middle, Haralick's detector implemented with Canny's filter. Note how the contour is broken at the junctions, due to the bad estimate of the gradient direction, and the high number of edge pieces. Right: detailed view of meaningful boundaries on the region. There are two level lines, each corresponding to an edge part. Also note the accuracy of level lines on the plate

can be expressed implicitly (which ease the possible topological changes of the active contours) and can use image statistics [34, 138]. This section does not concentrate on any active contour model in particular, but tries to compare a generic model with meaningful boundaries. Such a comparison has already been made by Desolneux, Moisan and Morel [51] for meaningful boundaries. Even though these boundaries are only contrast-based, they show that they are very close to active contours in general and particularly to the model of Kimmel and Bruckstein [95].

Let us briefly give a generic active contour model: it is a curve that fits shape contours (hence contrast should be large along the contour) and which is also as smooth as possible. This can be formulated in a variational way. An optimal curve minimizes an energy of the type

$$E(C) = \int_C g(|Du(C(s))|) + \lambda h(\mathrm{curv}(C(s)))\, ds, \tag{A.1}$$

where Du is the gradient of a given grey-level image, g is a nonincreasing function, $\mathrm{curv}(C(s))$ is the curvature of C at point $C(s)$, h is a nondecreasing function and s is the arc-length. The optimal curve is a trade-off between the external energy depending on the image gradient, and the internal energy depending on the curve itself only. Such a model can accurately give the position of the contour.

In [51], Desolneux, Moisan and Morel compared the MB model with variational snake theory. This may seem a bit weird since the MB model only uses contrast observations along a curve, while snakes are also required to be smooth. In fact, the explanation for natural images is that contrasted boundaries often locally coincide with objects. Thus, they are also incidentally smooth. Whereas smoothness seems to be optional for the detection, it may give a better localization of the contour. This section attempts to analyze the possible influence of smoothness in the detection. In particular, is smoothness fundamental in the detection? The result is the following: there are only few additional detections, while the position of the maximal meaningful boundaries may change a little bit. The NFA may significantly decrease. Moreover, detections using only contrast and only regularity yield comparable results. This proves, a contrario, that contrast and regularity are not independent in natural images.

An a contrario model of regularity has been proposed in [28]. It assumes that the variation of the orientation of the tangent between two samples is a random value uniformly distributed in $(-\pi, \pi)$. Thus, the implicit a contrario model is random walks with isotropic and independent increments. This model is not really adapted for the following reason. All the detected curves are level lines, thus boundaries of compact sets. As a consequence, they do not self-intersect. While the local influence is not clearly visible, this implies that long level lines are much more regular than random walks. This logically leads to an overdetection of long level lines because the independence assumption is strongly violated at very long range. The proposed solution is to stick to Helmholtz principle: "no detection in white noise". Thus, the regularity of level lines in white noise has to be learned, and be used as the *a priori* distribution.

Definition of local regularity

Let $l_0 > 0$ be a fixed positive value. Let C be a rectifiable planar curve, parameterized by its length. Let $x = C(s_0) \in C$. With no loss of generality, it can be assumed that $s_0 = 0$.

DEFINITION A.1 *The regularity of C at x (at scale l_0) is the quantity*

$$R_{l_0}(x) = \frac{\max(|x - C(-l_0)|, |x - C(l_0)|)}{l_0}. \tag{A.2}$$

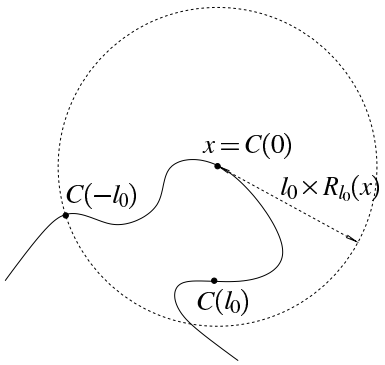


Figure A.2: Regularity definition. The regularity at x is obtained by comparing the radius of the circle with l_0. The radius is equal to l_0 if and only if the curve is a straight line. If the curve has a large curvature, the radius will be small compared to l_0

Of course, this definition really makes sense if the length of C is larger than $2l_0$. This definition of regularity (see Fig. A.2) is related to the Hausdorff dimension of C around x. First, $R_{l_0}(x) \leqslant 1$, with equality if and only if either $C((-l_0, 0))$ or $C((0, l_0))$ is a line segment. On the contrary, if $R_{l_0}(x)$ is small, then the curve is highly curved around x.

The regularity $R_{l_0}(x)$ can also be viewed as a function of the local curvature. Indeed, if C is a circle with large enough radius ρ, then

$$R_{l_0}(x) = \operatorname{sinc}\left(\frac{l_0}{2\rho}\right), \text{ where } \operatorname{sinc} x = \frac{\sin x}{x}. \tag{A.3}$$

This approximation is valid when l_0 is small compared to ρ. In this case, the regularity is a nonincreasing function of the curvature.
This definition is not purely local, but it is also less sensitive to noise compared to differential measures as the curvature. Let

$$\mathcal{H}_{l_0}(r) = P(x \in C,\ C \text{ is a white noise level line and } R_{l_0}(x) > r). \tag{A.4}$$

This distribution only depends on l_0 and can be empirically estimated. Of course, it is learned on level lines whose length is much larger than l_0 in order to avoid quantization effects. *Remark:* As expected, the distribution

$\mathcal{H}_{l_0}$ is very different in white noise and natural images. In natural images, the histogram of R_{l_0} has a peak at 1, corresponding to real objects boundaries (which often contain alignments). In some textured images, such as paintings, most edges are not real but subjective and this is clearly visible on the histogram of R_{l_0}. See Fig. A.3. The distribution also clearly depends on l_0. When l_0 grows, the histogram mode moves to lower values. However, the qualitative behavior is the same as above. In Sect. A.1.2, these distributions are used to numerically approximate the Hausdorff dimension of white noise level lines. As expected, they are found to be much smoother than (self-intersecting) isotropic random walks.

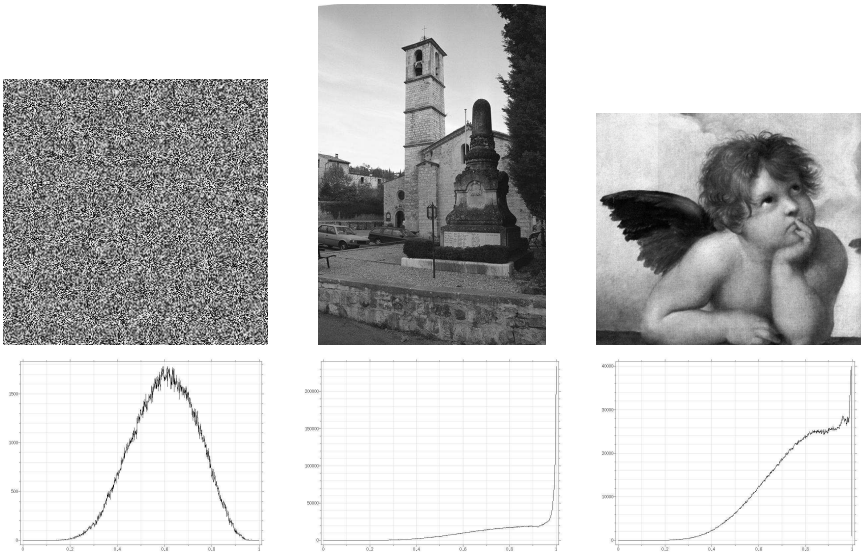


Figure A.3: Regularity histograms. Upper row: a white noise image, a scanned photograph and a scanned photograph of a painting. Bottom row: the three regularity histograms for $l_0 = 10$. Since its histogram vanishes near 1, white noise does not contain any alignments or smooth curves, as foreseen. Nearly all natural images (containing true edges) have a regularity histogram like the second one. The third image contains mostly subjective edges, as it is composed of painted strokes. As a consequence, the regularity histogram is much less concentrated around 1 as for "natural" images. After unzooming the three images (with an adequate smoothing before downsampling), the first histogram remains unchanged (scale invariance), while the other two have regularity histograms like the second one. Indeed, after unzooming, most textures and small scale features disappear, and small gaps get filled

Meaningful contrasted and smooth boundaries

Now that a background model of regularity is at hand, detecting regular curves a contrario is possible. It is natural to assume that contrast and regularity are independent in the background model. Thus

$$P(C \text{ is contrasted and smooth}) = P(C \text{ is smooth}) \times P(C \text{ is contrasted}).$$

DEFINITION A.2 *Let C be a level line. Let*

$$\nu = \min\{|Du(x)|,\, x \in C\}, \tag{A.5}$$

$$\rho = \min\{|R_l(x)|,\, x \in C\}, \tag{A.6}$$

be respectively the minimal contrast and regularity along C. Let

$$NFA_{cs}(C) = N_{ll} H_c(\nu)^{l/2} \left(\mathcal{H}_{l_0}(\rho)\right)^{l/2l_0}. \tag{A.7}$$

A curve C is an ε-meaningful smooth boundary if $NFA_{cs}(C) < \varepsilon$.

The number of false alarms is the product of number of level lines and the probability that the contrast and the regularity are simultaneously larger than the observed values along a curve with prescribed length taken in the background model. The probability is computed in the a contrario model where contrast and regularity are independent and local observations are mutually independent.
The choice of l_0 is a natural question. Of course l_0 should be larger than the Nyquist's distance. It should not be too large either. In the experiments, $l_0 = 10$. But, since NFAs are additive, several reasonable values of l_0 may be chosen (say $l_0 = 5$, 10, 20). It is then necessary to multiply the NFAs by the number of test values. In practice, changing l_0 influences the number of samples and best NFAs are attained for small l_0. All the refinements defined in Chap.3 can of course be applied here.

Experiments on smooth meaningful boundaries

In experiments, detection results are qualitatively equivalent with or without regularity. On the other hand, NFA may decrease a lot for smooth boundaries. Even though the detection remains unchanged in one single image, it is still interesting to decrease the NFA as much as possible. Indeed, the boundaries of a whole database (and not a single image) may be interesting. Any database clearly has a size much less than 10^{15}. Thus, curves with a NFA lower than 10^{-15} in a single image will be 1-meaningful in any database, and therefore allways detected. The regularity criterion does not qualitatively change the result. This is conform to the observation of Desolneux et al. in [51]. Remark also that adding the regularity criterion does not eliminate irregular level lines that were already detected thanks to contrast. Indeed,

$$NFA_{cs}(C) \leqslant N_{ll} H_c(\nu)^{l/2},$$

(with the same notations as in Def. A.2) since $\mathcal{H}_r(\rho) \leqslant 1$. This results in detecting more lines, which is the purpose: check whether or not there were misdetections because regularity was not taken into account. Of course, the NFA of smooth boundaries decreases a lot (about 10^{-15}), and this can slightly modify maximal meaningful boundaries.
An interesting experiment is to define a NFA for smooth boundaries, with no care of contrast, as

$$NFA_{reg}(C) = N_{ll} \left(\mathcal{H}_{l_0}(\rho)\right)^{l/2l_0}. \tag{A.8}$$

Let us apply this definition to the level lines of the image of Fig. A.4. Most edges in the desk image are retrieved with this definition. Pieces of objects boundaries coincide with pieces of level lines, and they can be detected either by regularity or contrast, or both.
In Fig. A.5, locally straight structures are also contrasted but the gradient distribution exhibits large values (since the texture variations are important). This explains why contrasted meaningful boundaries lose many lines. In this case, our local regularity criterion allows to characterize this elongated structures.

Figure A.4: Regularity detectability. The original is the left most. In middle, the 204 detected contrasted smooth boundaries as defined in Def. A.2. On the right, the 96 smooth boundaries, with no contrast information, defined in (A.8). All the main boundaries are already present. Of course, contrast may be the main cause of small NFA, since regularity acts at larger scales. For instance, the window panes have NFA about 10^{-150} with contrast and 10^{-15} with regularity only (which still make them detectable in an image database containing 10^{15} images, which is far larger than any existing database). The desk on the bottom right has a NFA equal to 10^{-60} with contrast and 10^{-20} with regularity, which is already very small

Comparison with active contours

The variational formulation allows active contours to give the optimal position of the curve very accurately. However, all the models also share some common drawbacks:

1. they assume that there is a contour: they cannot be used as a detection algorithm. This also explains why active contours are also introduced in Bayesian models, where the real question is: knowing that one object is present, what is the best candidate?

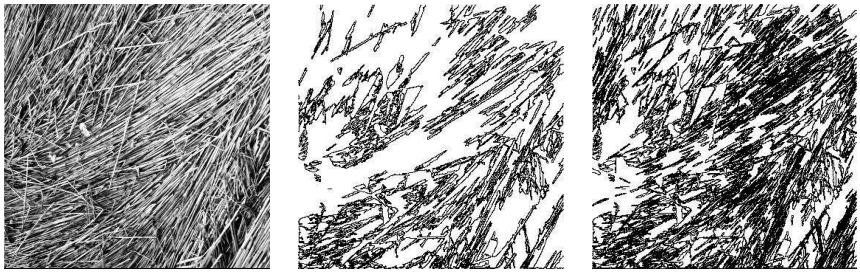


Figure A.5: Influence of regularity. On the left, the original texture contains a lot of elongated structure. Because the texture shows large contrast variations, meaningfulness is a very strict criterion and contrasted meaningful boundaries miss many details (middle). In this case, local regularity is important and smooth and contrasted boundaries allow to retrieve missing lines

2. Initialization is crucial.
3. The optimal balance parameter λ between the contrast and regularity term (which, for homogeneity reasons, can also be viewed as a scale parameter) is unknown and depends on the image. It has a strong influence on the result.

By only considering the homogeneity of the different energy terms, active contours minimize a potential of the form $Lg(|Du|) + \lambda Lh(\operatorname{curv} C)$, L being the length of the curve. Let us now go back to the meaningful smooth boundary model. For a meaningful curve, the quantity

$$(H_c(|Du|)^{L/2}\mathcal{H}_{l_0}(R_{l_0}(C)))^{L/2l_0}$$

is small. The logarithm of this probability leads to an expression of the type

$$L\left(E_{ext}(|Du|) + E_{int}(\operatorname{curv} C)\right),$$

where E_{ext} is a non increasing function of $|Du|$, and E_{int} is a non decreasing function of the curvature. The model is qualitatively alike a snake model. Nevertheless, there are three major differences:

1. In the meaningful boundaries method, there is a quantitative criterion to decide if the curve has to be detected. Contrarily to snakes algorithms, meaningful boundaries detection is *not* a minimization algorithm. It is well known in active contours models that the minimizer's energy value has no interpretation. All that can be said is that a candidate is better than another one. In a Bayesian setting, this relates to the fact that the *a* posteriori probability is only known up to a multiplicative factor. Our model gives a meaning to the energy-like term. Thus, there is no need for a minimization since there are thresholds under which a candidate has to be detected.
2. Meaningful boundaries are level lines. Thus, no initialization by hand is needed.
3. The weight functions g and h as well as the scale parameter λ are trivially defined in the model, with a clear meaning.

Numerical estimation of the Hausdorff dimension of a curve

Let us first recall the definition of the Hausdorff dimension of a set.

DEFINITION A.3 *[59] We say that $(B_i)_{i\in I}$ forms a covering of C if $C \subset \cup_{i\in I} B_i$. It is a δ-covering of C if moreover, for all $i \in I$, $|B_i| < \delta$, where $|B_i|$ is the diameter of B_i. The Hausdorff measure of dimension α of a set C is defined by*

$$\mathcal{H}^\alpha(C) = \lim_{\delta\to 0} \inf_{(B_i)\delta-covering} \sum_i |B_i|^\alpha. \tag{A.9}$$

The Hausdorff dimension of C is $\inf\{\alpha, \mathcal{H}^\alpha(C) = 0\}$.

The regularity measure in (A.2) allows to numerically estimate the Hausdorff dimension of a curve C. How to estimate the quantity on the right hand side of (A.9)? The problem is that it makes no sense to let $\delta \to 0$ for digital curves. Indeed, even for white noise, the precision is bounded from below by the Nyquist's distance. Let us assume that the curve is self-similar. This allows to examine it at larger and larger scales, instead of letting δ go to 0. Let us cut a curve with length $L = 2Nl$ in N chunks of length $2l$. The regularity $R_l(x_i)$ is measured at the middle point x_i of each piece. The balls with radius $R_l l$ nearly form a covering of C. It is not a covering because the endpoint of the curve chunk may not be the most remote point from the center (see Fig. (A.2)). Nevertheless, the measure of C can be approximated by

$$\mathcal{H}^\alpha(C) \simeq \sum_{i=1}^{N} (2lR_l(x_i))^\alpha \simeq 2^{\alpha-1} L l^{\alpha-1} \overline{R_l}^\alpha,$$

where $\overline{R_l}$ is the mean regularity along C. (The sum is approximated as the mean value summed $N = \frac{L}{2l}$ times.) Let us now consider the curve λC with $\lambda > 1$. The same procedure as above can be made with chunks whose length is equal to $2\lambda l$. This amounts to evaluate the measure of λC by

$$\mathcal{H}^\alpha(\lambda C) \simeq 2^{\alpha-1} \lambda L (\lambda l)^{\alpha-1} \overline{R_{\lambda l}}^\alpha.$$

But, if pieces of curves of length $2l$ are used instead, then

$$\mathcal{H}^\alpha(\lambda C) \simeq 2^{\alpha-1} \lambda L l^{\alpha-1} \overline{R_l}^\alpha.$$

Thus

$$\lambda^\alpha \overline{R_{\lambda l}}^\alpha = \lambda \overline{R_l}^\alpha,$$

yielding

$$\log(\overline{R_{\lambda l}}) = \left(\frac{1}{\alpha} - 1\right) \log \lambda + \log \overline{R_l}. \tag{A.10}$$

It is possible to evaluate α by examining the histograms of R_l as a function of l.
For random walks with independent increments, the experimental value is $\alpha = 2.02$, whereas the true dimension is 2. For level lines in white noise, the computed value is $\alpha = 1.78$. As expected, the level lines of a white noise image are more regular than random walks.

A.2 A reader digest in clustering analysis

Clustering methods have been and are still the object of applied and theoretical research in many different fields, such as statistical pattern recognition, data mining, image processing, biomedical sciences, etc. It is not the aim of this section to present a complete overview of clustering techniques, but just to provide enough information to justify why a particular technique can be preferred (there is no universal "best" clustering algorithm, and choices and compromises have to be made). A good review of clustering techniques by Jain *et al.*, from a statistical pattern recognition viewpoint, can be found in [89]. The main concepts can also be found in Duda and Hart [56], Hastie *et al.* [79] and Kaufman and Rousseeuw [93] textbooks.

Most of the clustering algorithms are either partitional, either hierarchical methods. While partitional methods produce a single partition, hierarchical methods produce a nested series of partitions. In this sense, they provide a totally different data description and should not be considered as two competiting techniques. However, as shall be seen, because of their different nature, the corresponding strategies for cluster validity assessment may be quite different.

A.2.1 Partitional clustering methods

Let us denote by $\mathcal{T} = \{T_k,\ k \in \{1,\dots,M\}\}$ the data set where each pattern T_k is a D-dimensional feature vector, and by $d_T : \mathcal{T} \times \mathcal{T} \to \mathbb{R}^+$ the dissimilarity measure. Assuming for the moment that the partition size c is given, the goal of a partitional clustering algorithm is to identify the partition $\mathcal{P}(\mathcal{T}) = \{\mathcal{T}_1,\dots,\mathcal{T}_c\}$ on $\mathcal{T}$ that optimizes a criterion function. Parametric methods as mixture decomposition will not be addressed here, since there is no a *a priori* knowledge on the underlying probability distribution. (In these methods, the data set is assumed to be drawn from a mixture of c underlying parametric distributions, and the goal is to determine the involved parameters; the standard algorithm is the Expectation-Maximization algorithm [44].) Hence, since there are approximately $c^M/c!$ ways of partitioning a set of M elements into c subsets (a Stirling number of the second kind), optimizing the criterion function by exhaustive search is intractable and iterative optimization procedures are needed.

The simplest and most widely used family of criteria function is the one of related minimum variance criteria [56, 93]. The energy to be minimized here is

$$E = \frac{1}{2}\sum_{m=1}^{c} n_m \langle d_m \rangle,$$

where n_m is the number of points in the m-th cluster, and

$$\langle d_m \rangle = \frac{1}{n_m^2}\sum_{T_i \in \mathcal{T}_m}\sum_{T_j \in \mathcal{T}_m} d_T(T_i, T_j)$$

is the average dissimilarity measure between points in the m-th cluster. If $\mathcal{T}$ was a subset of a vector space, and d_T was the squared Euclidean distance, the resulting criteria would be the sum of variances of each cluster

$$\sum_{m=1}^{c}\sum_{T \in \mathcal{T}_m} \|T - \langle T_m \rangle\|_2^2, \quad \text{where } \langle T_m \rangle = \frac{1}{n_m}\sum_{T \in \mathcal{T}_m} T.$$

Strictly speaking, this criterion only makes sense when clusters are isotropic, multivariate normally distributed. Moreover, the solution is not invariant to linear transformations of the data. Many variations on this method exists, taking any Minkowski metric or the squared Mahalanobis distance instead of the squared Euclidean distance [89]. Notice however that all these methods are based on the notions of medoid or centroid (barycenter) of a set of points and this does not make sense unless patterns live in a vector space.

Related minimum variance criteria suffer from the problem that partitions that split large clusters may be favored over ones that maintain the integrity of natural clusters [56]. When natural clusters have very different

number of points, the partition minimizing this criteria may not reveal the intrinsic structure of the data. Another weakness of these methods is the lack of ability to extract a very dense cluster embedded in the center of a diffuse cluster. Besides, the partition solution has to be found by iterative optimization procedures. These iterative procedures are to be initialized by a reasonable initial partition and solution can be trapped in local minima [89].

Other popular criterion functions, also defined only when patterns live in Euclidean (or Hermitian) spaces, and closely related to the previous ones, can be derived based on the "within cluster" scatter matrix $W\left(\mathcal{P}(\mathcal{T})\right)$, and the "between cluster" scatter matrix $B\left(\mathcal{P}(\mathcal{T})\right)$ [56],

$$
\begin{aligned}
W\left(\mathcal{P}(\mathcal{T})\right) &= \sum_{m=1}^{c} \sum_{T \in \mathcal{T}_m} \left(T - \langle T_m \rangle\right) \cdot \left(T - \langle T_m \rangle\right)^{\mathrm{T}}, \\
B\left(\mathcal{P}(\mathcal{T})\right) &= \sum_{m=1}^{c} n_m \left(\langle T_m \rangle - \langle T \rangle\right) \cdot \left(\langle T_m \rangle - \langle T \rangle\right)^{\mathrm{T}}, \\
S &= \sum_{m=1}^{c} \left(T - \langle T \rangle\right) \cdot \left(T - \langle T \rangle\right)^{\mathrm{T}} = W\left(\mathcal{P}(\mathcal{T})\right) + B\left(\mathcal{P}(\mathcal{T})\right),
\end{aligned}
$$

where $\langle T \rangle$ is the barycenter of all patterns in the data set, and S is the "total" scatter matrix, which is a constant given the data, independent on the partition. One can define optimal partitions as minimizers of $\mathrm{tr}\left[W\left(\mathcal{P}(\mathcal{T})\right)\right]$ (or equivalently maximizers of $\mathrm{tr}\left[B\left(\mathcal{P}(\mathcal{T})\right)\right]$); this turns out to be a minimum variance criterion. Another possiblity is to minimize $\det\left[W\left(\mathcal{P}(\mathcal{T})\right)\right]$, whose solution is invariant to linear transformations of the data. In any case, combinatorial optimization is intractable and one has to consider iterative procedures.

Iterative methods for partitional clustering Most partitional methods are based in the definition of c elements from the pattern space called centrotypes, each of them defined to be a representative object of one of the clusters. The criterion function to be minimized is usually the sum of the average dissimilarities between each centrotype and all the other patterns of the same cluster. Typically, iterative methods begin by initialising a set of c centrotypes; each pattern is then assigned to the cluster corresponding to its closest centrotype (for the considered dissimilarity measure), and centrotypes are re-computed in order to mimimize the criterion function. The iteration ends when centrotypes do not change. The computational efficiency of this approach depends on how easily centrotypes can be computed. The c-means algorithm [113] (also referred in the literature as k-means) runs typically in $O(M)$ [23]; indeed, in this algorithm the dissimilarity measure is the squared Euclidean distance and centrotypes are the clusters' barycenters, which can be easily computed using an update equation. A similar algorithm can be obtained by using the ℓ_1-norm as dissimilarity measure; the centrotypes for this measure (which is more robust to outliers than the squared Euclidean distance) are the clusters' medians.

When the dissimilarity measure does not lead to a closed form representation for the centrotypes, a method known as k-medoid which allows clustering with respect to any specified dissimilarity measure can be used [93]. In this method, centrotypes (the so called medoids) are restricted to be patterns from the data set, and as before patterns are assigned to the cluster corresponding to its closest centrotype. The goal is then to select, among all M patterns, the c centrotypes which minimize the sum of the average dissimilarities between each centrotype and all the other patterns of the same cluster. A widely used implementation for the k-medoid method is the Partitioning Around Medoids algorithm (PAM), by Kaufman and Rousseeuw [93]. PAM consists of two phases; in the first one, a method for selecting the initial set of c centrotypes or medoids is applied. The second phase is an iterative procedure, where in each iteration the set of centrotypes is updated by analyzing all possible pairs of patterns such that one pattern is a centrotype and the other is not, and by swaping the pair which most reduces the value of the criterion function. The cost of a single iteration is $O(c(M-c)^2)$.

A.2.2 Hierarchical clustering methods

While partitional clustering algorithms construct a single partition with c clusters (a "flat" description), hierarchical methods obtain a recursive structure. Since they represent data in different ways, partitional and hierarchical methods do not really compete with one another. Indeed, when data is to be described in terms of classes, subclasses, subsubclasses (e.g. a biological taxonomy), flat representations do not make sense, and hierarchical methods are needed. There are, of course, many applications in which data is not inherently hierarchical, and one has to make a choice among clustering methods from both types. Hierarchical methods are more versatile than partitional methods, and can deal with many differently shaped clusters, but generally they are more time consuming.

Depending on the direction they build the hierarchy, these clustering methods can be agglomerative (bottom-up) or divisive (top-down). The former, which are usually computationally simpler, start with each single point as a cluster, and iteratively merge the closest pair of clusters in the sense of a chosen dissimilarity measure. The generic algorithm is as follows [89]:

1. *Initialization*: compute the proximity matrix (the matrix containing the dissimilarity between each pair of patterns).
2. Find the most similar pair of clusters using the proximity matrix. Merge these two clusters.
3. Update the proximity matrix according to this merging.
4. Repeat steps 2 and 3 until all patterns are in one cluster.

At each iteration step, two clusters are merged. The procedure builds up a tree or dendrogram, where leaves are the M elements of $\mathcal{T}$ (step 1). At level l, there are $M - l$ nodes, each node being a cluster. At level $l + 1$, the closest clusters from level l are merged (step 2). By "closest", we mean the pair $\mathcal{T}_i$ and $\mathcal{T}_j$ minimizing a given distance or proximity measure $\delta(\mathcal{T}_i, \mathcal{T}_j)$ between clusters. Different strategies for updating the proximity matrix lead to different hierarchical clustering methods. (Moreover, since all these algorithms are merging methods, they admit a variational formulation and can be solved as an energy minimization problem; see [131], Chap. 3.) Lance and Williams [101] define a class of methods by specifying a generalized recurrence formula for updating the proximity matrix:

$$\delta(\mathcal{T}_i \cup \mathcal{T}_j, \mathcal{T}_k) = \alpha_i\, \delta(\mathcal{T}_i, \mathcal{T}_k) + \alpha_j\, \delta(\mathcal{T}_j, \mathcal{T}_k) + \beta\, \delta(\mathcal{T}_i, \mathcal{T}_j) + \gamma\, |\delta(\mathcal{T}_i, \mathcal{T}_k) - \delta(\mathcal{T}_j, \mathcal{T}_k)|,$$

where parameter values α_i, α_j, β and γ characterize the particular clustering method. Let us describe the most popular ones:

- Choosing $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ and $\gamma = -1/2$, leads to the following distance between clusters:

  $$\delta_{min}(\mathcal{T}_p, \mathcal{T}_q) = \min_{T_i \in \mathcal{T}_p, T_j \in \mathcal{T}_q} d_T(T_i, T_j).$$

  The corresponding algorithm is known as *single-linkage algorithm* [89, 56]. Here the nearest-neighbor points determine the nearest subsets. If elements in $\mathcal{T}$ are viewed as nodes of a graph, merging $\mathcal{T}_p$ and $\mathcal{T}_q$ corresponds to adding an edge between the nearest points in $\mathcal{T}_p$ and $\mathcal{T}_q$. This procedure generates a tree, and if one lets the procedure evolve up to having a single cluster containing all points, the result is a *minimal spanning tree*.

- Taking $\alpha_i = \alpha_j = \gamma = 1/2$, $\beta = 0$, yields

  $$\delta_{max}(\mathcal{T}_p, \mathcal{T}_q) = \max_{T_i \in \mathcal{T}_p, T_j \in \mathcal{T}_q} d_T(T_i, T_j).$$

  The resulting algorithm is called *complete-linkage algorithm* [89, 56]. Here distance between two clusters is given by the farthest pair of points in the two clusters. This procedure produces a graph in which edges

connect all of the nodes in a cluster. When the nearest clusters are merged, edges between every pair of nodes in the two clusters are added. If the diameter of a partition is defined as the largest diameter for clusters in the partition, then each iteration of the complete-linkage algorithm increases the diameter of the partition as little as possible.

- Taking $\alpha_i = n_i/(n_i+n_j)$, $\alpha_j = n_j/(n_i+n_j)$, and $\beta = \gamma = 0$, leads to a group average method, where

$$\delta_{avg}(\mathcal{T}_p, \mathcal{T}_q) = \frac{1}{n_p n_q} \sum_{T_i \in \mathcal{T}_p} \sum_{T_j \in \mathcal{T}_q} d_T(T_i, T_j).$$

- Some clustering methods based on barycenters, like Ward's minimum variance method [162], can also be represented in terms of Lance and Williams formula. For Ward's method, $\alpha_i = (n_i+n_k)/(n_i+n_j+n_k)$, $\alpha_j = (n_j+n_k)/(n_i+n_j+n_k)$, $\beta = -n_k/(n_i+n_j+n_k)$, $\gamma = 0$, and the corresponding cluster proximity measure is

$$\delta_{ward}(\mathcal{T}_p, \mathcal{T}_q) = \frac{n_p n_q}{n_p + n_q} \| \langle T_p \rangle - \langle T_q \rangle \|_2^2,$$

where $\langle T_p \rangle$ and $\langle T_q \rangle$ denote the barycenters of $\mathcal{T}_p$ and $\mathcal{T}_q$, respectively.

Time and memory complexity of algorithms given by Lance and Williams formula are studied in [41]. Overall, the time required for hierarchical clustering is $O(M^2 \log M)$, and the memory complexity is $O(M^2)$.

In practice, if clusters are compact and well separated, all methods yield the same results. However, when this is not the case, the resulting partitions may be quite different. Depending on the cluster proximity measure, different methods of clustering can be more or less successful with different types of clusters. The single-linkage algorithm suffers from the "chaining effect": a single corrupted point somewhere in between two compact clusters may lead to an unwanted merging between them [89, 56]. However, this property is very useful if one wants to detect elongated clusters.

The complete-linkage algorithm tends to produce compact clusters with small diameters. However, patterns assigned to a cluster can be much closer to patterns in other clusters [79, 56].

The single-linkage and the complete-linkage algorithms are both sensitive to outliers, since they rely on extremal measures. One way to reduce the influence of outliers is using δ_{avg} as cluster proximity measure, though the improvement is often not good enough. Besides, average methods have another drawback compared to single or complete linkage methods: they are not invariant under monotone transformations on the dissimilarity measure d_T (invariance of the former ones is a consequence of being based on extremal values) [79].

To end this section, let us make a few general remarks. In Sect. A.2.1, one of the main assumptions is that the number of clusters c was given, for partitional clustering algorithms. Then, the goal was to find the c-partition on the data optimizing a global criterion (in practice iterative methods are used, and the convergence to a global minimum is not ensured). Agglomerative hierachical clustering methods perform good in making local decisions about cluster merging, since they make use of the proximity matrix. As the hierarchy is built by means of local optimization, the level corresponding to a c-partition will not correspond in general to a global optimum (unless clusters are compact and well separated). For instance, Ward's method will not lead to the same c-partition than a c-means method, despite the fact that both attempt to minimize variance. In this sense, one would rather say that partitional methods are better than hierarchical methods. But how to be sure that there are exactly c groups of patterns in the data? Is the criterion function well adapted to the shape of clusters that are present in the data? From this viewpoint, hierarchical clustering may be more appealing than partitional ones. Another argument in favor for hierarchical clustering methods is their versatility and their ability to cope with differently shaped clusters. For instance, the single linkage algorithm can deal with non-isotropic, elongated or concentric clusters, while partitional methods like c-means can only deal with isotropic clusters. Since their outputs are nested series of partitions, ranging from M clusters to one single cluster, one can imagine methods to determine the number of clusters, as stopping rules of the merging process. If stopping rules are correctly designed, hierarchical methods would also be able to detect clusters having different densities or different number of points.

A.2.3 Cluster validity analysis and stopping rules

The great variety of clustering methods that have been proposed in the recent past has been followed by an increasing interest in clustering validation methods. In [70], a comprehensive study of these techniques is presented.
Cluster validity analysis deals with assessing the validity of classifications obtained from the application of clustering procedures. There are different validation approaches [54, 70], depending on the amount of prior information on the data. This section deals with *internal validation tests*, which consist in determining if the structure is intrinsically adapted to the data. In other words, internal tests are derived from some *internal criteria* measuring the suitability of the clustering structure for the original data set, with no other information than the data themselves.
Classical issues in cluster validity analysis are the assessment of individual cluster validity, and the assessment of a whole partition. (In some applications it can also be required to assess the validity of a dendrogram; this problem is not addressed here.) In what follows, these two issues are briefly summarized.

Partition validity assessment

A relevant question to address in order to assess the validity of a partition, is deriving the number of clusters [54], denoted by c. Notice that by solving this problem, it cannot be ensured that the c clusters are valid clusters. The most common approach to decide how many clusters are best consists in finding partitions for $c = 1, \ldots, c_{max}$ and optimizing a measure $G(c)$ of partition adequacy, which is usually based on the within-cluster and between-cluster variability. When applied to hierarchical clustering methods, these cluster validity assessment techniques are known as *global stopping rules*, because the choice of c can be seen as stopping the merging process (in the agglomerative case) at a certain level of the dendrogram.

When dealing with hierarchical classifications, another approach to determine the most appropriate number of clusters are *local stopping rules*. In the agglomerative case, these rules are *merging criteria* for deciding whether two clusters should be merged. Usually, the merging process is continued until it is decided, for the first time, that two clusters should not be aggregated.

Milligan and Cooper [121], and Dubes [54], present comparative studies of some stopping rules. Milligan and Cooper's paper provides a particularly comprehensive Monte-Carlo evaluation of these rules, by comparing 30 local and global stopping rules. In their simulation experiment, only strongly clustered data sets (internally cohesive and well separated clusters) were considered. Hence, since clustering this kind of data should not be a challenging problem, techniques that do not perform well on it are also expected to be inefficient when dealing with any data set. The main conclusion of this experiment is that only five or maybe six of the compared rules perform quite well on strongly clustered data. One can also observe that the majority of the stopping rules described in the study are based on heuristics and lack of theoretical foundation. Those derived from rigorous statistical techniques, assume in general hypotheses on the data which are unrealistic in most real applications (*e.g.* multivariate normal distribution for the patterns). In order to briefly illustrate the considered stopping rules, it is worth to describe Calinski and Harabasz's index [24] and Duda and Hart's rule [56], since these methods provided the best results.

- Calinski and Harabasz propose a *global stopping rule* for assessing partitions, by choosing the partition size c that maximizes the index

$$G(c) = \frac{\frac{1}{c-1}\mathrm{tr}\left[B\left(\mathcal{P}(\mathcal{T})\right)\right]}{\frac{1}{M-c}\mathrm{tr}\left[W\left(\mathcal{P}(\mathcal{T})\right)\right]},$$

  where $B\left(\mathcal{P}(\mathcal{T})\right)$ and $W\left(\mathcal{P}(\mathcal{T})\right)$ are, respectively, the between- and within-cluster scatter matrices of a c-partition $\mathcal{P}$, defined in section A.2.1. The index $G(c)$ is the ratio between the total within-cluster sum of squared distances about the centroids, and the total between-cluster sum of squared distances. This index is only defined for sets of patterns living in an Euclidean space. Moreover, since the index is based on the sum of squares criterion, it has a tendency to partition the data into hyperspherical shaped clusters, having roughly equal numbers of patterns [70] (this is probably the main reason for its first position in

Milligan and Cooper's ranking, since their data was strongly clustered, and clusters contained almost the same numbers of points and were pretty isotropic).

- Duda and Hart proposed the "$Je(2)/Je(1)$" *local stopping rule* for deciding whether or not a cluster should be splitted into two subclusters. The rule consists in computing the ratio between the total within sum of squared distances about the centroids of the two clusters for the two-cluster solution ($Je(2)$), and the within sum of squared distances about the centroid when only one cluster is present ($Je(1)$). The method is based in considering a null hypothesis, assuming that all patterns come from a normal distribution, whose mean and variances are empirically estimated over the whole data set. The null hypothesis of one single cluster is rejected if $Je(2)/Je(1)$ is smaller than a specified critical value, fixed by a significance level for the hypothesis testing. While considering a normal distribution as a null hypothesis and using the sum of squared distances may not be well adapted to real clustering problems (particularly when the number of patterns in the data set is not as large to be well represented by an asymptotic distribution), the proposed *a contrario* formulation is appealing from our point of view.

Let us finish the discussion on partition validity assessment by quoting one of Bock's conclusions from his work on significance tests in cluster analysis [22], where a comparison between global and local methods is made:

> *Some care is needed when applying any test for clustering, bearing in mind that different types of clusters may be present simultaneously in the data, and that the number of clusters is, in some sense, dependent on the intended level of information compression. Thus, a global application of a cluster test to a large or high-dimensional data set will not be advisable in most cases. However, a "local" application (...) to a specific part of the data will often be useful for providing evidence for or against a prospective clustering tendency.*

Validity assessment of individual clusters

The problem is now to decide, among the candidate clusters furnished by the clustering procedure, which are the ones that correspond to "natural" clusters. But what does a "natural" cluster look like? As pointed out by Gordon [70], it may be difficult to specify a relevant definition of ideal cluster for a particular data set. However, clusters must reveal structure in the data, and can be detected as opposite to a complete absence of structure. Thus, in order to decide whether the candidate clusters are significant, they can be compared with some appropriate random distribution. This leads to a general methodology for cluster validity analysis, based on the statistical approach of hypothesis testing [22, 69, 70]. Following Bock [22], this framework consists in:

1. Design a null hypothesis $\mathcal{H}$ for the absence of class structure in the data (a *background model*, or *null model*), meaning that patterns are sampled from a "homogeneous" population. Then, "heterogeneity" or "clustering structure" are involved in the alternative hypothesis $\mathcal{A}$.
2. Define a test statistic, which will be used as validity index to discriminate between $\mathcal{H}$ and $\mathcal{A}$.
3. If, for a given significance level (error probability) α, the test statistic of the observed data exceeds the corresponding critical value c_α, the null hypothesis $\mathcal{H}$ is rejected, in favor of $\mathcal{A}$.

This general framework can be adapted for assessing the validity of individual clusters. A general approach within this framework is Monte-Carlo validation, which is described in [70]. Assume one wants to assess the validity of an observed cluster $\mathcal{T}_i$ having n patterns, in a data set having M patterns. In the Monte-Carlo validation method, data sets of M patterns are simulated under the background model, and classified using the same clustering procedure that was used to classify the original data. The test statistic is computed for those clusters having n patterns, and the distribution of the test statistic is estimated. Then, using the value of the test statistic of $\mathcal{T}_i$, one can compute the significance level of rejecting $\mathcal{H}$. Two popular test statistics are the maximum F test and the U statistic (see Bock [22] and Gordon [70]).

The specification of appropriate null models for data is the subject of the study presented in [69]. These models, which specify the distribution of patterns in the absence of structure in the data, can be of two types:

- *Standard (data-independent) null models*. Two well known standard null models are the *Poisson model* and the *Unimodal model* [22]. The main problem with the Poisson model is the choice of the region R within which patterns are uniformly distributed (standard choices for normalized data are the unit hypercube and the unit hypersphere). The Unimodal model assumes that the joint distribution of the variables describing the patterns is unimodal, but the choice of the distribution may not be easy.
- *Data-influenced null models*. Here the data is used to influence the specification of the null model. Examples of these null models are the Poisson model where R is chosen to be the convex hull of the data set, or the *Ellipsoidal model*, which is a multivariate normal distribution, whose mean and covariance matrix are given by the data set.

In [69], Gordon concludes that the results of the tests considerably depend on the choice of the null model, and that, in general, the results based on data-influenced null models are more relevant than those obtained using a standard null model.

A.3 Three classical methods for object detection based on spatial coherence

This section addresses some issues of the generalized Hough transform [14], of which variations are probably the most widely used techniques in object detection. Two frequently used techniques for robust transformation estimation will also be described: geometric hashing[100, 170] and the RANSAC algorithm [61].

A.3.1 The generalized Hough transform

In [14], Ballard proposed a generalization to the Hough transform [81] allowing to detect arbitrary planar shapes undergoing similarity transformations. Most of object detection and recognition systems using transformations clustering are based on the generalized Hough transform. The basic idea is to quantize the transformation space into D-dimensional cells. Each transformation point T_i is quantized, and then votes for one of these cells. In practice, noise and image quantization induce localization errors in the extracted features, and one has to take into account uncertainty in computing T_i. Thus, each pairing of model and image features defines a volume of possible transformations, so it should cast a vote into each cell intersecting this volume (see [72] for an error analysis when using line segments as features).
As like as all techniques based on histograms in multidimensional spaces, the generalized Hough method is very sensitive to the choice of quantization precision (this remark also holds for Lamdan and Wolfson's Geometric Hashing [170, 100], described in Sect. A.3). Most of the time, the cell size is chosen by problem specific *ad hoc* arguments (see [112] for an example). However, in the general case, quantization effects may lead to several problems:

- Similar transformation points may vote for different cells. In order to reduce this problem, either votes are counted by adding the votes of neighboring cells (using a sliding window) in the case of no uncertainty in T_i, or, when uncertainty is considered, a vote is casted into each cell intersecting the uncertainty volume.

- In the plane similarity case, for instance, if one wants to do a fine discretization of the 4-D transformation space in order to perform accurate detection, the search space is too large for exhaustive search. Coarse to fine techniques applied to transformation clustering, first introduced by Stockman [156], can deal with this complexity problem, but there is no reason why the most voted cells at the finer scale correspond to the most voted ones at coarser scales.

- From the detection viewpoint, the cells size is also crucial. Indeed, if quantization is too fine, cells will not have enough votes and correct instances will be missed (false negatives). On the other side, choosing a very coarse quantization increases the likelihood of large clusters occurring at random (false positives). Moreover, as pointed out by Grimson and Huttenlocher in [72], using a local neighborhood or casting multiple votes, and reducing the number of cells to reduce the search space, both methods greatly increase the chance of large random clusters.

These remarks partially motivate our decision of using the clustering techniques described in Chapter 8, along with the validity assessment method proposed in the same chapter. Indeed, the proposed methodology does not suffer from quantization problems.

The generalized Hough transform is with geometric hashing [100, 168, 170], and the alignment method [85] one of the most popular voting schemes. Given two shapes, the geometric hashing method aims at determining if there is a transformed subset of the features from one shape, that matches a subset of the features of the other one. The aligment method is a similar voting method. The generalized Hough transform method, instead of voting over all possible configurations of shapes, consists in voting over all possible transformations mapping a shape to another one. Like for all techniques based on histograms in multidimensional spaces, these voting methods are very sensitive to the choice of quantization precision (too large bins may lead to false matches, and too small bins may produce misses). Besides, most of the time, the size of the hash table and the amount of parameters (size of the bins in the voting stage, threshold for the amount of votes in each bin, *etc.*) are crippling. The complexity of these voting schemes increases with the invariance degree; affine invariant shape retrieval

in large databases is intractable. All these properties make the local features not suitable for shape retrieval in large databases.

A.3.2 Geometric hashing

In order to illustrate the geometric hashing algorithm, let us present it in the case of similarity or affine transformations.

A query shape $\mathcal{S}$ is searched in a set of shapes

Preprocessing (off line). For each shape $\mathcal{S}'_i$ in the set of shapes:

1. Extract local invariant features from $\mathcal{S}'_i$. Assume n such features are found.
2. For each local basis b_j (*e.g.* a pair of points for similarity transformations, three non-collinear points for affine transformations) of features:
   (a) Compute the quantized coordinates (u, v) of all the remaining features, in the local basis.
   (b) Use the couple (u, v) as an index in a hash table, and write the information (i, b_j) in the corresponding bin (i is the index that identifies $\mathcal{S}'_i$).

Recognition stage (on line) For the query shape $\mathcal{S}$:

1. Extract local invariant features from $\mathcal{S}$. Assume n such features are found.
2. Choose arbitrarily a local basis (two or three points, depending on the considered invariance).
3. Compute the quantized coordinates (u, v) of all the remaining features, in the local basis.
4. For each of these coordinates, go to the corresponding bin in the hash table, and cast a vote for each pair (i, b_j) inscribed in the bin.
5. Keep only the pairs (i, b_j) which received more than a certain number of votes: each of this pairs stands for a potential match.
6. For each potential match, compute the best transformation (in the least squares sense) between all corresponding features, and check if the query features and the features from the corresponding shape, are well aligned. If not, go to (2) and choose another basis.

For affine invariant shape recognition, time complexity for the preprocessing stage is $O(n^4)$ for each shape in the set of shapes, and, if the access time to the hash table is $O(1)$, time complexity for the recognition stage is between $O(m)$ (when the first query basis chosen at random corresponds to a model in the set of shapes) and $O(m^4)$ (when no basis from the query shape corresponds to a model in the set of shapes).

A.3.3 A RANSAC based approach

The "RANdom SAmple Consensus" (RANSAC) algorithm by Fischler and Bolles [61], is certainly one of the most popular robust estimators in computer vision. It has proved very successful in stereo vision tasks, such as the estimation of homographies and fundamental matrices [78]. The main reasons of its success are its quite general nature, and its ability to deal with large proportions of outliers. Roughly speaking, in its general form, the RANSAC procedure to fit a model consists in randomly selecting a minimal subset of the data (*i.e.* a subset allowing to instantiate the model), then computing the number of inliers consistent with the instantiated model. These two steps are repeated for N minimal subsets of the data. The model having the largest number of inliers is chosen, and it is refined by re-estimating it from the corresponding set of inliers.
Our framework deals with M meaningful matches, and usually M is small enough to test for all corresponding similarity or affine transformations. Hence, using the same ideas, an elementary algorithm would be as follows:

1. For each element in the set of M pairs of local frames corresponding to meaningful matches:
   (a) Compute the associated transformation T.
   (b) Apply T to all query local frames, and compute their distances to their corresponding scene local frames.
   (c) Compute the number of inliers consistent with T, *i.e.* the pairs for which the distance is less than d pixels.
2. Choose the transformation T having the largest number of inliers.
3. Re-estimate T for all pairs of local frames determined as inliers (with a least squares method, for instance).

One can iterate this procedure on the set of outliers, in order to find other (less dominant) transformations. Even for this simple version of the algorithm, two problems arise: the choice of the distance threshold d, and the minimum number of inliers a model should have in order to be valid. The distance threshold d is usually chosen empirically; otherwise, it can be chosen by considering a significance level α, corresponding to the probability that a point is an inlier [78], what requires hypothetizing a model for the distribution of distances. Concerning the minimum number of inliers to assess model validity, generally it is also fixed by means of arbitrary rules. It seems reasonable to us that this minimum number of inliers depends on the distance threshold, but up to our knowledge, no effort has been done to establish this relation.

A.4 On the negative association of multinomial distributions

This section presents the notion of *negative association* (a strong notion of negative dependence) and summarize some relevant consequences, first reported by Joag-Dev and Proschan in [90]. Some proofs are also completed, when they were just oulined in the original paper. The result is then applied to multinomial distributions.

DEFINITION A.4 (NEGATIVE ASSOCIATION) *A set $\mathcal{X} = \{X_1, \ldots, X_n\}$ of real random variables is said to be negatively associated (NA) if for every two disjoint index sets $I, J \subset \{1, \ldots, n\}$,*

$$\mathbb{E}\left[f(X_i, i \in I) g(X_j, j \in J)\right] \leqslant \mathbb{E}\left[f(X_i, i \in I)\right] \cdot \mathbb{E}\left[g(X_j, j \in J)\right],$$

for all non-decreasing functions $f : \mathbb{R}^{\#I} \to \mathbb{R}$, $g : \mathbb{R}^{\#J} \to \mathbb{R}$ (a function $h : \mathbb{R}^k \to \mathbb{R}$ is said to be non-decreasing if $h(x_1, \ldots, x_k) \geqslant h(y_1, \ldots, y_k)$ whenever $x_1 \leqslant y_1, \ldots, x_k \leqslant y_k$).

Remark: Negative association is a natural generalization of negative correlation.

The negatively associated set $\mathcal{X} = \{X_1, \ldots, X_n\}$ verifies the following properties:
Property. For any non-decreasing functions f_i, $i \in \{1, \ldots, n\}$,

$$\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] \leqslant \prod_{i=1}^{n} \mathbb{E}\left[f_i(X_i)\right].$$

Proof: Define $f(x_1, \ldots, x_{n-1}) = \prod_{i=1}^{n-1} f_i(x_i)$ and $g(x_n) = f_n(x_n)$ for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Since f and g are both non-decreasing, it follows from Definition A.4 that

$$\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] \leqslant \mathbb{E}\left[\prod_{i=1}^{n-1} f_i(X_i)\right] \mathbb{E}\left[f_n(X_n)\right].$$

Using induction yields the desired result. □

Property. For all $(x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$\Pr\left(X_i \geqslant x_i \; \forall \; i \in \{1, \ldots, n\}\right) \leqslant \prod_{i=1}^{n} \Pr\left(X_i \geqslant x_i\right).$$

This follows immediately from Property A.4 for $f_i(x) = \chi_{[x \geqslant x_i]}$, the indicator function of event $[x \geqslant x_i]$. The following property is obvious from Definition A.4:
Property. Non-decreasing functions defined on disjoint subsets of a set of NA random variables are NA.

Property. The union of independent sets of NA random variables is NA.

Proof: Let $\mathbf{X}$ and $\mathbf{Y}$ be independent vectors such that for each one, its components are sets of NA random variables. Let $(\mathbf{X}_1, \mathbf{X}_2)$ and $(\mathbf{Y}_1, \mathbf{Y}_2)$ denote arbitrary partitions of $\mathbf{X}$ and $\mathbf{Y}$ respectively. Hence, the vector $(\mathbf{X}, \mathbf{Y})$ is NA if and only if $\mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1) g(\mathbf{X}_2, \mathbf{Y}_2)\right] \leqslant \mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1)\right] \mathbb{E}\left[g(\mathbf{X}_2, \mathbf{Y}_2)\right]$. Now,

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1) g(\mathbf{X}_2, \mathbf{Y}_2)\right] &= \mathbb{E}\left\{\mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1) g(\mathbf{X}_2, \mathbf{Y}_2) | \mathbf{Y}_1, \mathbf{Y}_2\right]\right\} \\
&= \sum_{(y_1, y_2)} \mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1) g(\mathbf{X}_2, \mathbf{Y}_2) | \mathbf{Y}_1 = y_1, \mathbf{Y}_2 = y_2\right] \\
&\qquad \cdot \Pr(\mathbf{Y}_1 = y1, \mathbf{Y}_2 = y2).
\end{aligned}$$

Since $(\mathbf{X}_1, \mathbf{X}_2)$ and $(\mathbf{Y}_1, \mathbf{Y}_2)$ are independent, $\{f(\mathbf{X}_1, \mathbf{Y}_1)|\mathbf{Y}_1 = y_1, \mathbf{Y}_2 = y_2\}$ and $\{g(\mathbf{X}_2, \mathbf{Y}_2)|\mathbf{Y}_1 = y_1, \mathbf{Y}_2 = y_2\}$ are parametric functions of random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively. Thus, because of the negative association of $\mathbf{X}$,

$$\mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1)g(\mathbf{X}_2, \mathbf{Y}_2)|\mathbf{Y}_1 = y_1, \mathbf{Y}_2 = y_2\right] \leqslant \mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1)|\mathbf{Y}_1 = y_1, \mathbf{Y}_2 = y_2\right] \mathbb{E}\left[g(\mathbf{X}_2, \mathbf{Y}_2)|\mathbf{Y}_1 = y_1, \mathbf{Y}_2 = y_2\right].$$

Hence,

$$\mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1)g(\mathbf{X}_2, \mathbf{Y}_2)\right] \leqslant \mathbb{E}\left\{\mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1)|\mathbf{Y}_1, \mathbf{Y}_2\right] \mathbb{E}\left[g(\mathbf{X}_2, \mathbf{Y}_2)|\mathbf{Y}_1, \mathbf{Y}_2\right]\right\}$$

Now, since the conditional expectations

$$\mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1)|\mathbf{Y}_1, \mathbf{Y}_2\right] \text{ and } \mathbb{E}\left[g(\mathbf{X}_2, \mathbf{Y}_2)|\mathbf{Y}_1, \mathbf{Y}_2\right]$$

are respectively $\mathbf{Y}_1$ and $\mathbf{Y}_2$ measurable functions, it follows that

$$\begin{aligned} h_1(\mathbf{Y}_1) &\equiv \mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1)|\mathbf{Y}_1, \mathbf{Y}_2\right] = \mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1)|\mathbf{Y}_1\right], \\ h_2(\mathbf{Y}_2) &\equiv \mathbb{E}\left[g(\mathbf{X}_2, \mathbf{Y}_2)|\mathbf{Y}_1, \mathbf{Y}_2\right] = \mathbb{E}\left[g(\mathbf{X}_2, \mathbf{Y}_2)|\mathbf{Y}_2\right]. \end{aligned}$$

Finally, using that $\mathbf{Y}$ is NA,

$$\begin{aligned} \mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1)g(\mathbf{X}_2, \mathbf{Y}_2)\right] &\leqslant \mathbb{E}\left[h_1(\mathbf{Y}_1)h_2(\mathbf{Y}_2)\right] \\ &\leqslant \mathbb{E}\left[h_1(\mathbf{Y}_1)\right] \mathbb{E}\left[h_2(\mathbf{Y}_1)\right] \\ &= \mathbb{E}\left[f(\mathbf{X}_1, \mathbf{Y}_1)\right] \mathbb{E}\left[g(\mathbf{X}_2, \mathbf{Y}_2)\right]. \qquad \square \end{aligned}$$

These results yield the following proposition.

PROPOSITION A.1 *A random vector $\mathbf{X} = (X_1, \ldots, X_n)$ having a multinomial distribution of index M and parameter $\mathbf{p} = (p_1, \ldots, p_n)$ (denoted by $\mathbf{X} \sim Mult(M, \mathbf{p})$), is NA.*

Proof: The variable $\mathbf{X}$ can be decomposed as

$$\mathbf{X} = \sum_{k=1}^{M} \mathbf{Y}_k,$$

where each $\mathbf{Y}_k \sim Mult(1, \mathbf{p})$, and the $\mathbf{Y}_k$'s are mutually independent. Since, for all $k \in \{1, \ldots, M\}$, all elements in $\mathbf{Y}_k$ are zero except for one whose value is 1, vector $\mathbf{Y}_k$ is NA. Indeed, for all I, J disjoint subsets of $\{1, \ldots, n\}$, for all non-decreasing functions $f : \mathbb{R}^{\#I} \to \mathbb{R}$, $g : \mathbb{R}^{\#J} \to \mathbb{R}$,

$$\begin{aligned} &\mathbb{E}\left[f(\mathbf{Y}_{k,i}, i \in I)g(\mathbf{Y}_{k,j}, j \in J)\right] \leqslant \mathbb{E}\left[f(\mathbf{Y}_{k,i}, i \in I)\right] \cdot \mathbb{E}\left[g(\mathbf{Y}_{k,j}, j \in J)\right] \\ &\quad \Leftrightarrow \mathbb{E}\left[\left(f(\mathbf{Y}_{k,i}, i \in I) - f(0, \ldots, 0)\right)\left(g(\mathbf{Y}_{k,j}, j \in J) - g(0, \ldots, 0)\right)\right] \\ &\quad \leqslant \mathbb{E}\left[f(\mathbf{Y}_{k,i}, i \in I) - f(0, \ldots, 0)\right] \cdot \mathbb{E}\left[g(\mathbf{Y}_{k,j}, j \in J) - g(0, \ldots, 0)\right]. \end{aligned}$$

The last inequality is true: the right member is non-negative because $f(\mathbf{Y}_{k,i}, i \in I) - f(0, \ldots, 0)$ and $g(\mathbf{Y}_{k,j}, j \in J) - g(0, \ldots, 0)$ are non-negative, and the left member is zero since $(f(\mathbf{Y}_{k,i}, i \in I) - f(0, \ldots, 0)$ and $g(\mathbf{Y}_{k,j}, j \in J) - g(0, \ldots, 0)$ cannot be non-zero at the same time.
Then, using Property A.4, it follows that $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}$ is NA. Finally, for all $l \in \{1, \ldots, n\}$, $X_l = \sum_{k=1}^{M} \mathbf{Y}_{k,l}$ are non-decreasing functions defined on disjoint subsets of $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_M\}$. This proves that $\mathbf{X}$ is NA (Property A.4). $\square$

Remark: Applying Property A.4 to the random vector $\mathbf{X}$ proves lemma 8.3, stated in section 8.3.

Index

Index

Bibliography

[1] S. Abbasi and F. Mokhtarian. Retrieval of similar shapes under affine transformation. In *Proceedings of International Conference on Visual Information Systems*, pages 566–574, Amsterdam, The Netherlands, 1999.

[2] A. Almansa, A. Desolneux, and S. Vamech. Vanishing point detection without any a priori information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):502–507, 2003.

[3] H. Alt and L. Guibas. Discrete geometric shapes: Matching, interpolation, and approximation. In J.-R. Sack and J. Urrutia, editors, *Handbook of Computational Geometry*, pages 121–153. Elsevier Science Publishers, 1999.

[4] H. Alt, C. Knauer, and C. Wenk. Matching polygonal curves with respect to the Fréchet distance. In *Proceedings of the 18th International Symposium on Theoretical Aspects of Computer Science*, pages 63–74, Dresden, Germany, February 15-17 2001.

[5] L. Alvarez, F. Guichard, P.-L. Lions, and J.-M. Morel. Axioms and fundamental equations of image processing: Multiscale analysis and P.D.E. *Archive for Rational Mechanics and Analysis*, 16(9):200–257, 1993.

[6] L. Alvarez, L. Mazorra, and F. Santana. Geometric invariant shape representations using morphological multiscale analysis and applications to shape representation. *Journal of Mathematical Imaging and Vision*, 18(2):145–168, 2002.

[7] S. Angenent, G. Sapiro, and A. Tannenbaum. On the affine heat flow for nonconvex curves. *Journal of the American Mathematical Society*, 1998.

[8] N. Ansari and E. J. Delp. Partial shape recognition: A landmark-based approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):470–483, 1990.

[9] N. Arnaud, F. Cavalier, M. Davier, and P. Hello. Detection of gravitational wave bursts by interferometric detectors. *Physical review D*, 59(8):082002–1 – 082002–9, 1999.

[10] H. Asada and M. Brady. The curvature primal sketch. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(1):2–14, 1986.

[11] K. Åström. Affine and projective normalization of planar curves and regions. In *Proceedings of European Conference on Computer Vision*, volume 2, pages 439–448, Stockholm, Sweden, 1994.

[12] K. Åström. Fundamental limitations on projective invariants of planar curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):77–81, 1995.

[13] F. Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183–193, 1954.

[14] D.H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[15] C. Ballester, V. Caselles, and P. Monasse. The tree of shapes of an image. *ESAIM: Control, Optimisation and Calculus of Variations*, 9:1–18, 2003.

[16] G. Barles and P.M. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4:271–283, 1991.

[17] R. Basri, L. Costa, D. Geiger, and D. Jacobs. Determining the similarity of deformable shapes. *Vision Research*, 38(15-16):2365–2385, 1998.

[18] M. Faisal Beg, M.I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Internation Journal of Computer Vision*, 61(2):139–157, 2005.

[19] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color- and texture-based image segmentation using the Expectation-Maximization algorithm and its application to content-based image retrieval. In *Proceedings of the International Conference on Computer Vision*, pages 675–682, Mumbai, India, 1998.

[20] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):509–522, 2002.

[21] E. Bienenstock, S. Geman, and D. Potter. Compositionality, MDL priors, and object recognition. In M.C. Mozera, M.I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*. MIT Press, 1998.

[22] H.H. Bock. On some significance tests in cluster analysis. *Journal of Classification*, 2:77–108, 1985.

[23] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In G. Tesario and D. Touretzky, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 585–592, Denver, Colorado, USA, 1995.

[24] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in statistics*, 3(1):1–27, 1974.

[25] J. Canny. A variational approach to edge detection. In *National Conference on Artificial Intelligence*, pages 54–58, Washington DC, USA, 1983.

[26] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[27] F. Cao. *Geometric Curve Evolution and Image Processing*, volume 1805 of *Lecture Notes in Mathematics*. Springer Verlag, 2003.

[28] F. Cao. Good continuations in digital images. In *Proceeding of International Conference on Computer Vision*, volume 1, pages 440–447, Nice, France, 2003.

[29] F. Cao. Application of the Gestalt principles to the detection of good continuations and corners in image level lines. *Computing and Visualisation in Science*, 7(1):3–13, 2004.

[30] F. Cao and L. Moisan. A geometrical scheme for curve evolution driven by curvature. *SIAM Journal of Numerical Analysis*, 39(2):624–646, 2000.

[31] F. Cao, P. Musé, and F. Sur. Extracting meaningful curves from images. *Journal of Mathematical Imaging and Vision*, 22(2-3):159–181, 2005.

[32] V. Caselles, B. Coll, and J.-M. Morel. A Kanizsa program. *Progress in Nonlinear Differential Equations and their Applications*, 25:35–55, 1996.

[33] V. Caselles, B. Coll, and J.-M. Morel. Topographic maps and local contrast changes in natural images. *International Journal of Computer Vision*, 33(1):5–27, 1999.

[34] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[35] P.B. Chapple, D.C. Bertilone, R.S. Caprari, and G.N. Newsam. Stochastic model-based processing for detection of small targets in non-gaussian natural imagery. *IEEE Transactions on Image Processing*, 10(4):554–564, 2001.

[36] F.S. Cohen, Z. Huang, and Z. Yang. Invariant matching and identification of curves using B-splines curve representation. *IEEE Transactions on Image Processing*, 4(1):1–10, 1995.

[37] T. Cohignac. *Reconnaissance de formes planes*. PhD thesis, Ceremade, Université Paris IX Dauphine, 1994.

[38] T. Cohignac, C. Lopez, and J.M. Morel. Integral and local affine invariant parameter and application to shape recognition. In *International Conference on Pattern Recognition*, pages A:164–168, 1994.

[39] J. Cortadellas, J. Amat, and F. de la Torre. Robust normalization of silhouettes for recognition applications. *Pattern Recogn. Letters*, 25(5):591–601, 2004.

[40] J.L. Cox and D.B. Karron. Digital Morse theory. Manuscript available at http://www.casi.net/D.DMT/D.Overview/AcademicPressPaper14-03, 1998.

[41] W.H.E. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24, 1984.

[42] F. de la Torre and M. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1-2-3):117–142, 2003.

[43] I. Debled-Rennesson, J.-L. Rémy, and J. Rouyer-Degli. Segmentation of discrete curves into fuzzy segments. Technical Report 4989, INRIA, 2003.

[44] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[45] R. Deriche. Using Canny's criteria to derive a recursively implemented optimal edge detector. *International Journal of Computer Vision*, 1(2):167–187, 1987.

[46] R. Deriche and O. Faugeras. Tracking line segments. *Image and Vision Computing*, 8(4):261–270, 1990.

[47] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful alignments. *International Journal of Computer Vision*, 40(1):7–23, 2000.

[48] A. Desolneux, L. Moisan, and J.-M. Morel. Edge detection by Helmholtz principle. *Journal of Mathematical Imaging and Vision*, 14(3):271–284, 2001.

[49] A. Desolneux, L. Moisan, and J.-M. Morel. Computational gestalts and perception thresholds. *Journal of Physiology - Paris*, 97(2-3):311–322, 2003.

[50] A. Desolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):508–513, 2003.

[51] A. Desolneux, L. Moisan, and J.-M. Morel. Variational snake theory. In S. Osher and N. Paragios, editors, *Geometric Level Set Methods in Imaging, Vision, and Graphics*. Springer Verlag, 2003.

[52] P.A. Devijver and J. Kittler. *Pattern recognition - A statistical approach*. Prentice Hall, 1982.

[53] I. Dryden. General shape and registration analysis. Technical report, University of Leeds, Department of Statistics, 1996.

[54] R . C. Dubes. How many clusters are best? – an experiment. *Pattern Recognition*, 20(6):645–663, 1987.

[55] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[56] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley and Sons, 2000.

[57] S.A. Dudani, K.J. Breeding, and R.B. McGhee. Aircraft identification by moment invariants. *IEEE Transactions on Computers*, 26(1):39–46, 1977.

[58] G. Dudek and J.K. Tsotsos. Shape representation and recognition from multiscale curvature. *Computer Vision and Image Understanding*, 2(68):170–189, 1997.

[59] L.C. Evans and R. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.

[60] O. Faugeras and R. Keriven. Some recent results on the projective evolution of 2d curves. In *Proceedings of IEEE International Conference on Image Processing*, volume 3, pages 13–16, Washington DC, USA, 1995.

[61] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the Association for Computing Machinery*, 24(6):381–395, 1981.

[62] M.A. Fischler and R.C. Bolles. Perceptual organization and curve partitioning. *IEEE Transactions on pattern analysis and machine intelligence*, 8(1):100–105, 1986.

[63] P. Frosini and C. Landi. Size theory as a topological tool for computer vision. *Pattern Recognition and Image Analysis*, 9(4):596–603, 1999.

[64] P. Frosini and C. Landi. Size functions and formal series. *Applicable Algebra in Engineering, Communication and Computing*, 12:327–349, 2001.

[65] Y. Gdalyahu and D. Weinshall. Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1312–1328, 1999.

[66] S. Geman and D. McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 52:5–21, 1987.

[67] G. Giraudon. Chaînage efficace de contour. Technical Report 0605, INRIA, 1987.

[68] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1989.

[69] A.D. Gordon. Null models in cluster validation. In W. Gaul and D. Pfeifer, editors, *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis, and Knowledge Organization*, pages 32–44. Springer Verlag, 1996.

[70] A.D. Gordon. *Classification*. Monographs on Statistics and Applied Probability 82, Chapman & Hall, 1999.

[71] Y. Gousseau. Comparaison de la composition de deux images, et application a la recherche automatique. In *proceedings of GRETSI 2003*, Paris, France, 2003.

[72] W.E.L. Grimson and D.P. Huttenlocher. On the sensitivity of the Hough transform for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(3):255–274, 1990.

[73] W.E.L. Grimson and D.P. Huttenlocher. On the verification of hypothesized matches in model-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(12):1201–1213, 1991.

[74] L. Guigues. *Modèles multi-échelles pour la segmentation d'images*. PhD thesis, Université de Cergy-Pontoise, 2003.

[75] G. Guy and G. Medioni. Inferring global perceptual contours from local features. *International Journal on Computer Vision*, 20(1):113–133, 1996.

[76] R. Haralick. Digital step edges from zero crossing of second directional derivatives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(1):58–68, 1984.

[77] C.G. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conference, Manchester*, pages 189–192, 1988.

[78] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[79] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2001.

[80] H. Helmholtz. *Treatise on Physiological Optics*. Dover, New York, 1962 (first published in 1867).

[81] P.V.C. Hough. *Methods and means for recognizing complex patterns*, 1962. U.S. Patent 3,069,654.

[82] M.K. Hu. Visual pattern recognition by moments invariants. *IRE Transactions on Information Theory*, 8:179–187, 1962.

[83] Z. Huang and F.S. Cohen. Affine-invariant B-spline moments for curve matching. *IEEE Transactions on Image Processing*, 5(10):1473–1480, 1996.

[84] D.P. Huttenlocher, G. Klanderman, and W. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[85] D.P. Huttenlocher and S. Ullman. Object recognition using alignment. In *International Conference of Computer Vision*, pages 267–291, London, UK, 1987.

[86] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[87] J. Illingworth and J. Kittler. A survey of the Hough transform. *Computer Vision, Graphics, and Image Processing*, 44(1):87–116, 1988.

[88] D.W. Jacobs. Robust and efficient detection of salient convex groups. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):23–37, 1996.

[89] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[90] K. Joag-Dev and F. Proschan. Negative association of random variables, with applications. *Annals of Statistics*, 11(1):286–295, 1983.

[91] G. Kanizsa. *Organization in Vision: Essays on Gestalt Perception*. Praeger, 1979.

[92] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1987.

[93] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, 1990.

[94] M.I. Khalil and M.M. Bayoumi. A dyadic wavelet affine invariant function for 2d shape recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), 2001.

[95] R. Kimmel and A.M. Bruckstein. On regularized Laplacian zero crossings and other optimal edge integrators. *International Journal of Computer Vision*, 53(3):225–243, 2003.

[96] J.J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.

[97] G. Koepfler and L. Moisan. Geometric multiscale representation of numerical images. In *Second International Conference on Scale Space Theories in Computer Vision*, volume 1682 of *Lecture Notes in Computer Science*, pages 339–350. Springer-Verlag, 1999.

[98] A. Krzyzak, S.Y. Leung, and C.Y. Suen. Reconstruction of two-dimensional patterns from Fourier descriptors. *Machine Vision and Applications*, 2:123–140, 1989.

[99] Y. Lamdan, J.T. Schwartz, and H.J. Wolfson. Object recognition by affine invariant matching. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 335–344, Ann Arbor, Michigan, U.S.A., 1988.

[100] Y. Lamdan and H.J. Wolfson. Geometric hashing: a general and efficient model-based recognition scheme. In *Proceedings of IEEE International Conference on Computer Vision*, pages 238–249, Tampa, Florida, USA, 1988.

[101] G.N. Lance and W.T. Williams. A general theory of classificatory sorting strategies. i. hierarchical systems. *Computer Journal*, 9:373–370, 1967.

[102] C.C. Lin and R. Chellappa. Classification of partial 2-d shapes using Fourier descriptors. In *CVPR86*, pages 344–350, 1986.

[103] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1998.

[104] M. Lindenbaum. An integrated model for evaluating the amount of data required for reliable recognition. *IEEE Trans. Pattern Analysis Machine Intelligence*, 19(11):1251–1264, 1997.

[105] M. Lindenbaum and A. Bruckstein. On recursive, $O(N)$ partitioning of a digitized curve into digital straight segments. *Transactions on Pattern Analysis and Machine Intelligence*, 15(9), 1993.

[106] J.L. Lisani. *Shape Based Automatic Images Comparison*. PhD thesis, Université Paris 9 Dauphine, France, 2001.

[107] J.L. Lisani, L. Moisan, P. Monasse, and J.-M. Morel. On the theory of planar shape. *SIAM Multiscale Modeling and Simulation*, 1(1):1–24, 2003.

[108] J.L. Lisani, P. Monasse, and L. Rudin. Fast shape extraction and applications. Technical Report 2001-16, CMLA, ENS Cachan, 2001.

[109] S. Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, 1998.

[110] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[111] D.G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publisher, 1985.

[112] D.G Lowe. Object recognition from local scale-invariant features. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, 1999.

[113] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability*, volume 1, pages 63–74, 1967.

[114] D. Marr. *Vision*. Freeman Publishers, 1982.

[115] D. Marr and E. Hildreth. Theory of edge detection. *Proceeding of the Royal Society of London*, B-207:187–207, 1980.

[116] G. Matheron. *Random Sets and Integral Geometry*. John Wiley and Sons, 1975.

[117] W. Metzger. *Gesetze des Sehens*. Waldemar Kramer, 1975.

[118] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[119] M.I. Miller, A. Trouvé, and L. Younes. On the metrics and Euler-Lagrange equations of computational anatomy. *Annual Review of Biomedical Engineering*, 4:375–405, 2002.

[120] M.I. Miller, A. Trouvé, and L. Younes. Geodesic shooting for computational anatomy. To appear, 2003.

[121] G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[122] L. Moisan. Affine plane curve evolution: A fully consistent scheme. *IEEE Transactions on Image Processing*, 7(3):411–420, 1998.

[123] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *International Journal of Computer Vision*, 57(3):201–218, 2004.

[124] F. Mokhtarian. Silhouette-based isolated object recognition through curvature scale space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):539–544, 1995.

[125] F. Mokhtarian, S. Abbasi, and J. Kittler. Efficient and robust retrieval by shape content through curvature scale space. In *Proceedings of International Workshop on Image Databases and MultiMedia Search*, pages 35–42, Amsterdam, The Netherlands, 1996.

[126] F. Mokhtarian, S. Abbasi, and J. Kittler. Robust and efficient shape indexing through curvature scale space. In *Proceedings of British Machine Vision Conference*, pages 53–62, Edinburgh, UK, 1996.

[127] F. Mokhtarian and A.K. Mackworth. A theory of multiscale, curvature-based shape representation for planar curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):789–805, 1992.

[128] P. Monasse. Contrast invariant image registration. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume 6, pages 3221–3224, Phoenix, Arizona, USA, 1999.

[129] P. Monasse. *Représentation morphologique d'images numériques et application au recalage, Morphological Representation of Digital Images and Application to Registration*. PhD thesis, Université Paris 9 Dauphine, France, 2000.

[130] P. Monasse and F. Guichard. Fast computation of a contrast invariant image representation. *IEEE Transactions on Image Processing*, 9(5):860–872, 2000.

[131] J.-M. Morel and S. Solimini. *Variational Methods in Image Segmentation*. Birkhauser, 1995.

[132] P. Musé, F. Sur, and J.-M. Morel. Sur les seuils de reconnaissance des formes. *Traitement du Signal*, 20(3):279–294, 2003.

[133] M. Niethammer, S. Betelu, G. Sapiro, A. Tannenbaum, and P.J. Giblin. Area-based medial axis of planar curves. *International Journal of Computer Vision*, 60(3):203–224, 2004.

[134] S. Obdržálek and J. Matas. Local affine frames for image retrieval. In *CIVR'02: Proceedings of International Conference The Challenge of Image and Video Retrieval*, pages 318–327. Springer-Verlag, 2002.

[135] C. Olson and D.P. Huttenlocher. Automatic target recognition by matching oriented edge pixels. *IEEE Transactions on Image Processing*, 6(12):103–113, 1997.

[136] C. Orrite and J.E. Herrero. Shape matching of partially occluded curves invariant under projective transformation. *Computer Vision and Image Understing*, 93(1):34–64, 2004.

[137] C. Orrite and J.E. Herrero. Shape matching of partially occluded curves invariant under projective transformation. *Computer Vision and Image Understanding*, 93:34–64, 2004.

[138] N. Paragios and R. Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *International Journal of Computer Vision*, 46(3):223–247, 2002.

[139] X. Pennec. Toward a generic framework for recognition based on uncertain geometric features. *Videre: Journal of Computer Vision Research*, 1(2):58–87, 1998.

[140] E. Persoon and K.S. Fu. Shape discrimination using Fourier descriptors. *SMC*, 7(3):170–179, 1977.

[141] H.V. Poor. *An Introduction to Signal Detection and Estimation*. Springer Texts in Electrical Engineering. Springer Verlag, 2nd edition, 1994.

[142] C.A. Rothwell. *Object Recognition Through Invariant Indexing*. Oxford Science Publications, 1995.

[143] C.A. Rothwell, A. Zisserman, D.A. Forsyth, and J.L. Mundy. Planar object recognition using projective shape representation. *International Journal of Computer Vision*, 16:57–99, 1995.

[144] E. Rubin. *Visuell wahrgenommene Figuren*. Copenhagen, Gyldendals, 1921.

[145] P. Salembier and J. Serra. Flat zones filtering, connected operators, and filters by reconstruction. *IEEE Transactions on Image Processing*, 4(8):1153–1160, 1995.

[146] G. Sapiro and A. Tannenbaum. Affine invariant scale-space. *International Journal of Computer Vision*, 11(1):25–44, 1993.

[147] J. Sato and R. Cipolla. Quasi-invariant parameterisations and matching of curves in images. *International Journal of Computer Vision*, 28(2):117–138, 1998.

[148] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[149] S. Sclaroff and A. Pentland. Modal matching for correspondence and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(6):545–561, 1995.

[150] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, 1982.

[151] D. Shen and H.H.S. Ip. Discriminative wavelet shape descriptors for recognition of 2-d patterns. *Pattern Recognition*, 32(8):151–165, 1999.

[152] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Ninth IEEE International Conference on Computer Vision (ICCV 203)*, pages 1470–1477, 2003.

[153] J. Sklansky and V. Gonzalez. Fast polygonal approximation of digitized curves. *Pattern Recognition*, 12:327–331, 1980.

[154] C.G. Small. *The Statistical Theory of Shapes*. Springer Verlag, 1996.

[155] C.V. Stewart. MINPRAN: a new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):925–938, 1995.

[156] G. Stockman, S. Kopstein, and S. Benett. Matching images to models for registration and object detection via clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(3):229–241, 1982.

[157] P.N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[158] R. Veltkamp and M. Hagedoorn. State-of-the-art in shape matching. In M.S. Lew, editor, *Principles of Visual Information Retrieval*, volume 19. Springer Verlag, 2001.

[159] R. Veltkamp and M. Tanase. Content-based image retrieval systems: A survey. Technical Report UU-CS-2000-34, Utrecht University, 2000.

[160] R.C. Veltkamp. Shape matching: similarity measures and algorithms. In *Proceedings of International Conference on Shape Modeling and Applications*, pages 188–197, Genova, Italy, 2001.

[161] C. C. Venters and M. D. Cooper. A review of content-based image retrieval systems. Technical Report jtap-054, University of Manchester, UK, 2000.

[162] J. H. Jr. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(2):236–244, 1963.

[163] G.H. Watson and S.K. Watson. Detection of unusual events in intermittent non-gaussian images using multiresolution background models. *Optical Engineering*, 35(11):3159–3171, 1996.

[164] I. Weiss. Noise-resistant invariants of curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):943–948, 1993.

[165] M. Wertheimer. Untersuchungen zur Lehre der Gestalt, II. *Psychologische Forschung*, 4:301–350, 1923. Translation published as Laws of Organization in Perceptual Forms, in Ellis, W. (1938). A source book of Gestalt psychology (pp. 71-88). Routledge & Kegan Paul.

[166] A. Winter and C. Nastar. Differential feature distribution maps for image segmentation and region queries in image databases. In *CBAIVL Workshop at Conference on Computer Vision and Pattern Recognition*, Fort Collins, Colorado, USA, 1999.

[167] A.P. Witkin. Scale space filtering. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1019–1021, Karlsruhe, Germany, 1983.

[168] H.J. Wolfson. Model-based object recognition by Geometric Hashing. In *Proceedings of the European Conference on Computer Vision*, pages 526–536, Antibes, France, 1990. Lecture Notes in Computer Vision 427, Springer Verlag.

[169] H.J. Wolfson. On curve matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):483–489, 1990.

[170] H.J. Wolfson and I. Rigoutsos. Geometric hashing: an overview. *IEEE Computational Science & Engineering*, 4(4):10–21, 1997.

[171] C.T. Zahn and R.Z. Roskies. Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, C-21(3):269–281, 1972.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399